MANGALORE UNIVERSITY

ADVANCED PREDICTIVE MODELS FOR AIR POLLUTION ANALYSIS IN SMART CITY SYSTEM

FINAL SYNOPSIS SUBMITTED
TO
MANGALORE UNIVERSITY
FOR THE AWARD OF THE DEGREE OF DOCTOR OF PHILOSOPHY
IN COMPUTER SCIENCE

BY

HARISHKUMAR
GADRESEARCH SCHOLAR

UNDER THE GUIDANCE OF
Dr. MANJAIAH D.H
PROFESSOR

DEPARTMENT OF POST-GRADUATE STUDIES AND RESEARCH IN COMPUTER SCIENCE
MANGALORE UNIVERSITY
MANGALAGANGOTRI-574 199
KARNATAKA, INDIA
DECEMBER-2021

Title: **ADVANCED PREDICTIVE MODELS FOR AIR POLLUTION ANALYSIS IN SMART CITY SYSTEM**

Name of Candidate: **HARISHKUMAR**

Name of Guide: **Prof. Manjaiah D.H**

Date of Registration: 15-03-2017

Place of Work: Department of Post-Graduate Studies and Research in Computer Science, Faculty of Science and Technology, Mangalore University, Mangalagangothri-574199,Mangalore-Karnataka-India

Chapter 1

Introduction and Literature Survey

1.1 Introduction

According to the World Health Organization (WHO) in the worldwide, more than 5.5 million people expire in every year as a result of air infection. Many of these deaths arise in more cities, where fatigue from factories, cars, and power stations fills the air with hazardous molecules. In response to the growing burden about the effects of air pollution, more cities develop their efforts to measure air pollution using the most recent technology, the Internet of Things (IoT), networks of connected sensors that congregate and send data. After enough data are collected and organized, machine learning techniques can be integrated into decision-making systems to make the strategic decisions on real-time data. Using the strategic decisions, cities can be mapped the areas of large pollution, track changes over time, identify polluters, and analyze possible interventions[1].

Data generated from IoT environment in the smart city is generally huge, spatio-temporal, different scales in numeric and heterogeneous characteristics. Analysis of these data requires data models for organizing and aggregating data, statistical and advanced machine learning models for the extraction of actionable insights from such data models. As the predictive models are reliable and better predictive performance on numerical data, development of high-performance predictive models for the analysis of data at different scales is challenging research issues in particular, for air pollution analysis of the emerging smart city systems [2].

Air pollution arises when injurious substances, along with particulates and biological molecules, are imported into Earth's atmosphere. It may cause disorder, allergies or death of mankind, it may

also cause harm to alternative living organisms such as creatures and food product and may damage the natural or assembled environment. Human activity and natural processes can both generate toxic waste that makes air infection. A toxic waste can be of a usual ancestor or artificial. Breeze impurity is categorized as essential or auxiliary. Essential poisons are commonly issued beginning of a mechanic, specific because dust as of an impulsive explosion, particulate matter, $PM_{2.5}$ from the climatic experimental condition, carbon monoxide (CO) gases starting engine automobile tire out fumes or the sulfur dioxide (SO_2) liberated beginning manufacturing plants. Auxiliary poisons are neither breathed straightforwardly or maybe, they shape in the discussion when essential poisons perform or act together. The earth stage breath is an outstanding case of an auxiliary toxin [3]. Outstanding of the boom for mechanical, populace and frequency of activity is developing quickly, individuals are confronting a think condition of for the discussing contamination. In a few nations, their relentless discuss contamination in the metropolis, not as it aggressive the well being of inhabitants other than the plant moreover the climate. Deepest locales, particulate matter $PM_{2.5}$ some absorption really on meteorological conditions, implying that weather alter might contain noteworthy impacts on particulate matter $PM_{2.5}$. All things considered, so distant small prove to be able to finish respecting the relationships of the $PM_{2.5}$ towards the weather varies. The ambiguous relationship turns around the troubles too in section the complication for the dependency of a different ingredient of a climatic factors, and to some extent, the active data on the effects of different components (e.g person variables) about for Particulate Matter $PM_{2.5}$ discuss value issue become more over the feasible to the investigation [4].

Recently the analysts of the National Aeronautics and Space Administration (NASA) addressed tended to that because of the air contamination, the consolidation of the $PM_{2.5}$ and dihydrogen monoxide steam lying on the western Pacific compose to create a solid tropical tornado the that loaded with dihydrogen monoxide steam, much all the more, developing the quality of tempests. In this manner, air contamination reaches towards the one of the fundamental sense that tempests wound up more grounded than previous in north america in current years [5]. A keenly interactive town is a metropolitan improvement view to incorporate the ICT and the IoT innovation in a secured pattern to deal with a cites benefits. These resources incorporate regional department's data structure, institutions, Athenaeum, transport frameworks, clinics, control stations, dihydrogen monoxide systems, waste management systems, rule prosecution, with extra social welfare [6]. A shrewd town is pushing toward utilizing metropolitan information processing and robotics towards progress the

expertise of skill of administrations. Information and communication technology allows urban authorities to collaborate specifically by for society & the metropolitan framework with to supervise what is going on into metropolitan, by means of town is emerging, and what to empower an improved aspect of the activity. Over the benefit of sensor collaborated through continuous checking frameworks, information is gathered starting the civilian & instruments after that prepared & scrutinized. The data as well as information accumulated be critical to handling wastefulness. [7].

1.2 Smart Cities

The concept of the smart city is getting to be more prevalent and realizable day by day. As the population in urban zones is expanding quickly, the city framework is extending consistently as well. According to United Countries, the urban populace is higher than the rural populace. Urban regions are domestic to 54 % of the world's population, concurring to an overview of the year 2014. Whereas only 30 percent of the World population was utilized to live in urban zones in 1950, it is anticipated that the urban population will go up to 66 percent within the coming years. With the fast and consistent development in the urban population, cities are facing financial, natural, and social dangers and concerns such as unemployment, environmental contamination, and social exclusion.

The thought of the smart city has risen as an endeavor to relieve the current issues related to urban areas and make superior livable cities. As the term, 'smart' is getting to be the center of discussion among analysts, approaches towards smart cities are getting to be more extensive amid different disciplines such as engineering, design, planning, and legislative issues [8]. The smart city concept is commonly related to information and communications technology (ICT) frameworks. It could be a concept that's improved by high-technology to supply an association between people data and the city so as to extend life quality and construct an economical, attractive greener city with competitive and pioneering commerce as shown in the figure 1.1.

1.3 Air pollution In Smart City

Air allows our living to inhale, it's the combination of gases that fill the environment, offering life to the plants and creatures that make Earth a particularly dynamic spot. Extensively talking, the

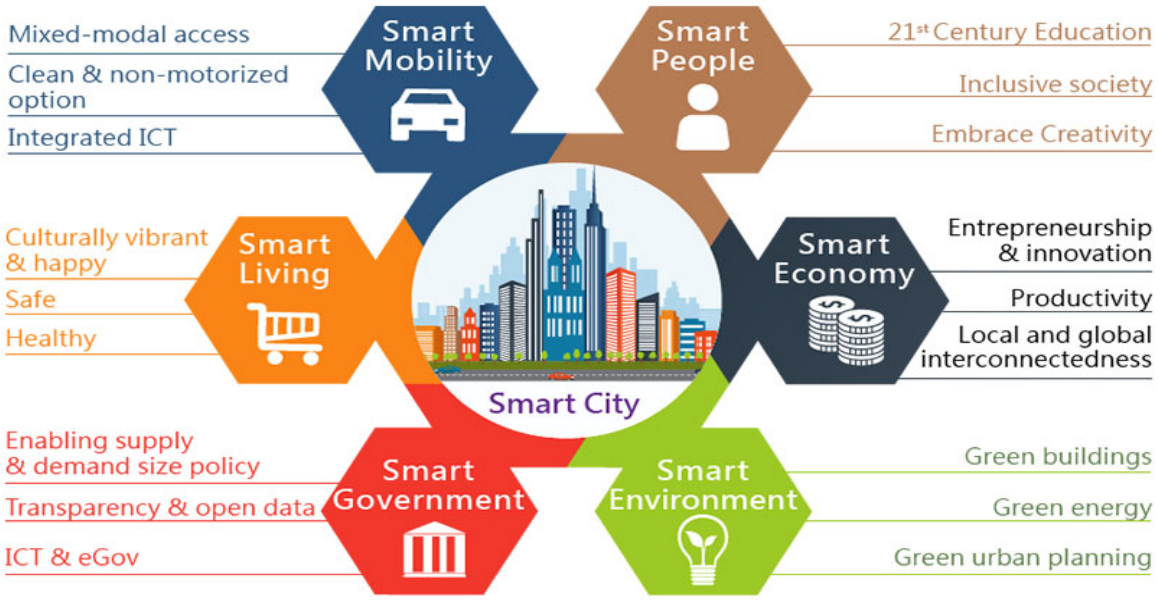

FIGURE 1.1: Smart City

air is for the most part comprised of two gases (78% Nitrogen and 21% Oxygen), with a couple of different gases (like carbon dioxide and argon) present in totally minute particles amounts. Air contamination is a gas (or a fluid or strong scattered through standard air) delivered in a large enough amount to hurt the soundness of individuals or different creatures, kill plants or prevent them from developing appropriately, harm or disturb some other part of the climate, (for example, making structures disintegrate), or cause some other sort of irritation [9].

1.3.1 Sources of Air Pollution

Transport sources consist of aircraft, ships, trains, boats, motor vehicles, and different off-road automobiles. Large-scale stationary fuel combustion sources include fossil fuel-burned institutional boilers, industrialized, electricity-generating plants, and home space ovens [10]. Miscellaneous sources consist of crop production flaming, jungle fires, structural fires and, coal refuse and a different nuclear solvent use. The following section illustrates the origin of air pollution and the literature on the latest analysis and prediction of air infection in the smart city system.

- **The Burning of Fossil Fuels**:

The greater part of the air contamination takes put because of the consumption of non-renewable energy sources like coal, oil, fuel to make essentials for transportation. The release of carbon monoxide at an undeniable level shows how much petroleum product is burned.

- **Industrial Emission**:

  Industrial activities exude a couple of toxins inside the air that influences the air quality more than ready to for sure envision. Particulate matter 2.5 and 10, Nitrogen dioxide, Sulfur dioxide, and carbon monoxide are key harms that are transmitted from industries that use coal and wood as their essential energy hot-spot for the age of their product [11].

- **Indoor Air pollution**:

  The use of poisonous things too called Volatile Organic Compounds (VOCs), lacking ventilation, uneven temperature, and dampness level can cause indoor examine tainting, regardless of whether you're inside the workplace, school, or at your agreeable home. House air defilement can take put because of ignorant factors, for an event, smoking tobacco inside a room or leaving mold-contaminated dividers untreated [12].

- **Agricultural Activities**:

  Agricultural activities have genuinely affected the diminishing air quality. In the first place, pesticides and manures are the fundamental sources to debase the incorporating air. Nowadays, pesticides and composts are mixed with unused meddlesome species which are not found in nature, for the expedient improvement of yields and vegetation [13].

- **Transportation** There is a thing like partly a billion automobiles on the roadway present-day single for each of two individuals in prosperous nations, for example, the US. Practically all of the authority are fueled by diesel & petrol motors that consume gas to discharge quality [14].

1.4 Literature Review

There are large numbers of attempts have been made by different researchers to forecast air pollution accurately using various techniques as shown in Table 1.1.

TABLE 1.1: Recent papers from standard publications are surveyed, reviewed, and summarized in the table below.

Begin of Table							
Ref No.	**year**	**Authors**	**Title**	**DataSet**	**Attributes**	**Methods**	**Performance**
[15]	2018	Huang, Keyong, et al.	Predicting monthly high-resolution $PM_{2.5}$ concentrations with random forest model in the north china plain	China	Particulate matter ($PM_{2.5}$)	Random-forest	The overall model performance R^2 and RPE is 0.88 and 18.7& respectively.
[16]	2018	Ung, Chau-Ren, Bing-Fang Hwang, and Wei-Ting Chen	Incorporating long-term satellite-based aerosol optical depth, localized land use data, and meteorological variables to estimate ground-level $PM_{2.5}$ concentrations in Taiwan from 2005 to 2015	Taiwan	Particulate matter($PM_{2.5}$)	Ten-fold cross validation	Based on the new satellite data developing the estimation model for inter-comparison studies and development of surface $PM_{2.5}$.

Continuation of Table 1.1

Ref No.	year	Authors	Title	DataSet	Attributes	Methods	Performance
[17]	2019	Stafoggia, Massimo, et al.	Estimation of daily PM_{10} and $PM_{2.5}$ concentrations in Italy, 2013–2015, using a spatio-temporal land-use random-forest model	Italian institute for environmental protection and research (ISPRA) Italy	PM_{10} and $PM_{2.5}$	Random-forest model	The predictions of daily and annual PM variability from an investigation of short-term and long-term health effects for cross-validation (CV), R^2 are 0.86 and 0.84 for PM_{10} and $PM_{2.5}$, respectively.
[18]	2020	Awan, Faraz Malik, Roberto Minerva, and Noel Crespi.	Improving road traffic forecasting using air pollution and atmospheric data: Experiments based on LSTM Recurrent Neural Networks	Madrid, Spain	CO, NO, NO_2, NO_x, and O_3	long short-term memory-recurrent neural network (LSTM-RNN)	The LSTM-RNN delivered values MAE is 0.105 and MSE is 0.017 for traffic sensor stream sensor "4303", and another traffic sensor stream sensor is "10387" is MAE is 0.136, and MSE is 0.029.
[19]	2020	Koo, Jian Wei, et al.	Prediction of air pollution index in kualalumpur using fuzzy time series and statistical models	Kualalumpur, Malaysia	Air Pollution Index (API)	ARIMA,TBATS, and fuzzy logic time series model	Developing the accurate forecasting model using fuzzy time series model gives better performance with MAPE and RMSE is 4.364% and 1.4704 respectively.

Continuation of Table 1.1

Ref No.	year	Authors	Title	DataSet	Attributes	Methods	Performance
[20]	2020	Shafii, Nor Hayati Binti, et al.	Forecasting of air pollution index $PM_{2.5}$ Using Support Vector Machine (SVM)	Malaysia	Air pollution index (API) of Particulate matter($PM_{2.5}$)	Support Vector Machine (SVM)	The forecast model is based on the API Model with the performance for Kota Bharu, and Tanah Merah stations RMSE, MAE 0.0625 and 0.0386 and RMSE, MAE 0.0589 and 0.0385 respectively.
[21]	2020	Chang, Yue-Shan, et al.	An ensemble learning-based hybrid model and framework for air pollution forecasting	Taiwan	Particulate Matter (PM_{10}, and $PM_{2.5}$)	GBTR, SVR, LSTM, and Hybrid model's	The prediction air pollution hybrid model shows the best results compare to the existing models.
[22]	2020	Liu, Duen-Ren, et al.	Air pollution forecasting based on attention-based LSTM neural network and ensemble learning	EPA Central Weather Bureau (CWB)	Particulate matter ($PM_{2.5}$)	MLP, SVR, XGBoost, LSTM, WALSTM, WALSTME-W	The proposed model (WALSTME-W) air pollution gives better results for accuracy in both short-term and long-term predictions.

Continuation of Table 1.1

Ref No.	year	Authors	Title	DataSet	Attributes	Methods	Performance
[23]	2020	Alyousifi, Yousif, et al.	Markov weighted fuzzy time-series model based on an optimum partition method for forecasting air pollution	Malaysia	Air pollution index (API)	Novel Markov weighted fuzzy time-series model	Air pollution forecasting, the proposed model has produced a superior forecasting accuracy compared to the conventional and advanced time-series method.
[24]	2020	Xu, Yinan, Hui Liu, and Zhu Duan.	A novel hybrid model for multi-step daily AQI forecasting driven by air pollution big data	China National Environmental Monitoring Centre (CNEMC)	AQI, $PM_{2.5}$, PM_{10}, SO_2, NO_2, O_2, and CO	VMD, LASSO, SAE, and Deep echo state network (DESN)	The proposed model gives better performance compared to the existing models and each component in the model gives a better contribution to the API forecasting.

End of Table

1.5 Objectives of The Proposed Research

The objectives of the proposed research work on air pollution analysis in smart city system:

1. Designing a data warehouse for air pollution data set statistical data analysis.
2. Visualization of framework and correlation analysis for air pollution dataset.
3. Developing high accuracy predictive models applicable for air pollution prediction in smart cities.
4. Developing a clustering model for Spatio-temporal data analysis for quality air prediction in smart cities.

1.6 Thesis Organization

The thesis consists of eight chapters and is organized as follows. It starts from Chapter 1 that gives an introduction and literature survey to the research work for air pollution prediction in smart cities. This includes a air pollution problem, survey of the literature studied in order to understand the problem statement, and presenting the objectives of the research.

- Chapter 2: This chapter introduces the research methodology and data organization. Moreover, this chapter introduces primary data and secondary data, then discuss the types of air quality monitoring stations in Taiwan country, in this research work that are essential for air pollution analysis. Furthermore, the framework followed in this research work is described the different data mining models and artificial neural network,and discuss the evaluation performance metrics to evaluate the proposed prediction and clustering performance models.
- Chapter 3: This chapter introduces the dataset used in this research work. Moreover, this chapter introduces the multidimensional data model in this research work that are essential for air pollution analysis. Furthermore, the framework followed in this research work is described the data acquisition and organization of the TAQMN air pollution data and the proposed the model for multidimensional data for air pollution analysis using snowflake schema.

- Chapter 4: This chapter presents a visualization of frame work and correlation analysis for air pollution data, using cross-correlation method. The threshold guidance and significance of testing of coefficients further implement the similarity analysis and causality detection. The correlation matrix for all air pollutants with different stations.

- Chapter 5: This chapter proposed the development of time series analysis for prediction of particulate matter ($PM_{2.5}$) using SARIMA Model. This new model considers the combination of auto regressive (AR), moving average MA(q), auto regressive moving average (ARMA(p,q)), auto regressive integrated moving average (ARIMA(p,d,q)), and seasonal ARIMA (SARIMA (p, d, q), (P, D, Q)).

- Chapter 6: This Chapter proposed the forecasting air pollution particulate matter ($PM_{2.5}$) using machine learning models gradient boosting regression model achieves excellent results in forecasting air pollution particulate matter ($PM_{2.5}$). When it is compared with the other machine learning models as the results of the GBR model achieves 99.8% future forecasting air pollution task.

- Chapter 7: This chapter present clustering analysis for air pollution using a deep learning model. This new model considers the combination of autoencoder, spatial autoencoder, and decoder techniques.

- Chapter 8: This chapter presents a summary of the thesis, including the conclusions we have drawn from our research. We display the scientific contributions that this research work reveals. Finally, we suggest several research directions ideas as a future work.

1.7 Conclusion

This chapter presents a background to the research problem with respect to air pollution forecasting and prediction of particulate matter ($PM_{2.5}$). An intensive literature survey has been carried out and the research gaps that led us to carry out this research work are incorporated. The motivation and identified research objectives are presented.

Chapter 2

Research Methodology and Data Organization

2.1 Introduction

The chapter introduces the dataset to be used in this research work. Research methodology is defined as an efficient method of gathering, organizing, incorporating, and data interpreting to improve the fundamental information about the marvel or article being scrutinized. This research data science approach is being proposed for exploring data mining models on advanced data models.

Data Science approach is being proposed for developing and deploying advanced data Mining Models for building relationships between input data and unknown in order to distinguish human activity recognition. Data science is a multidisciplinary field of statistics, applied mathematics,computer science and visualization for deep knowledge discovery through data inference and exploration[25]. This discipline regularly includes utilizing numerical and algorithmic methods to address probably the most systematically complex business issues, utilizing stores of raw data to sort out secret knowledge that lies underneath the surface as displayed in the figure 2.1.

It focuses on proof based logical thoroughness and building vigorous choice abilities [26]. A typical data science approach for building the data mining models is being followed and restructured in the proposed research and is shown in the shown in figure 2.2.

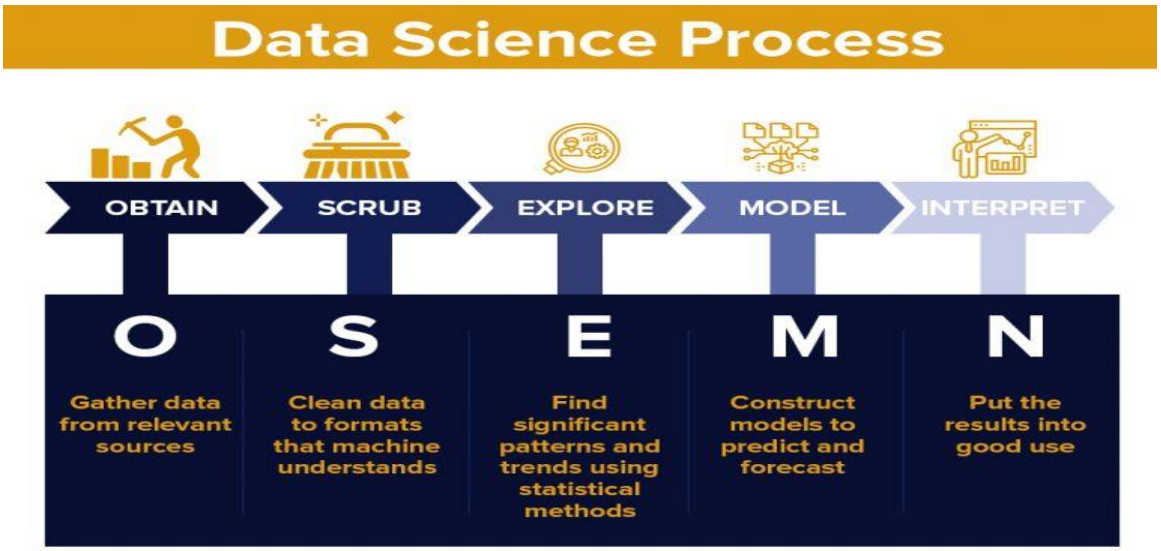

FIGURE 2.1: A typical building blocks of advanced data science process.

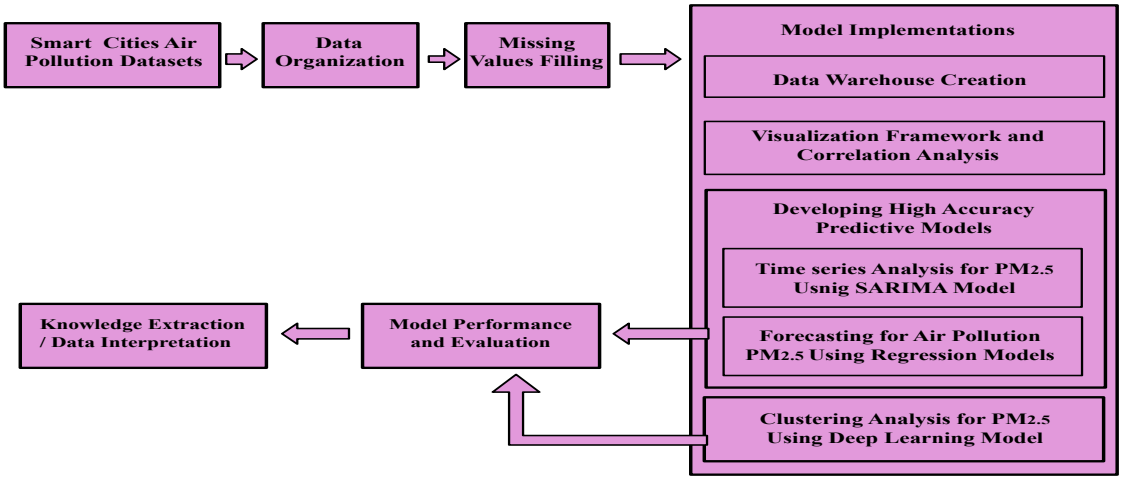
Smart Cities Air Pollution Datasets
Data Organization
Missing Values Filling
Model Implementations
Data Warehouse Creation
Visualization Framework and Correlation Analysis
Developing High Accuracy Predictive Models
Time series Analysis for PM2.5 Usnig SARIMA Model
Forecasting for Air Pollution PM2.5 Using Regression Models
Clustering Analysis for PM2.5 Using Deep Learning Model
Model Performance and Evaluation
Knowledge Extraction / Data Interpretation

FIGURE 2.2: A proposed model far with data science approach for building advanced data mining models.

2.2 Data acquisition and Organization

2.2.1 Primary Data

Air quality checking is a vital premise to advance air quality assurance and avoidance. In arrange to viably test into air quality, we must work and keep up an observing framework on a long-term premise to obtain high-quality, comprehensive, agent, and solid checking data. Taiwan country is

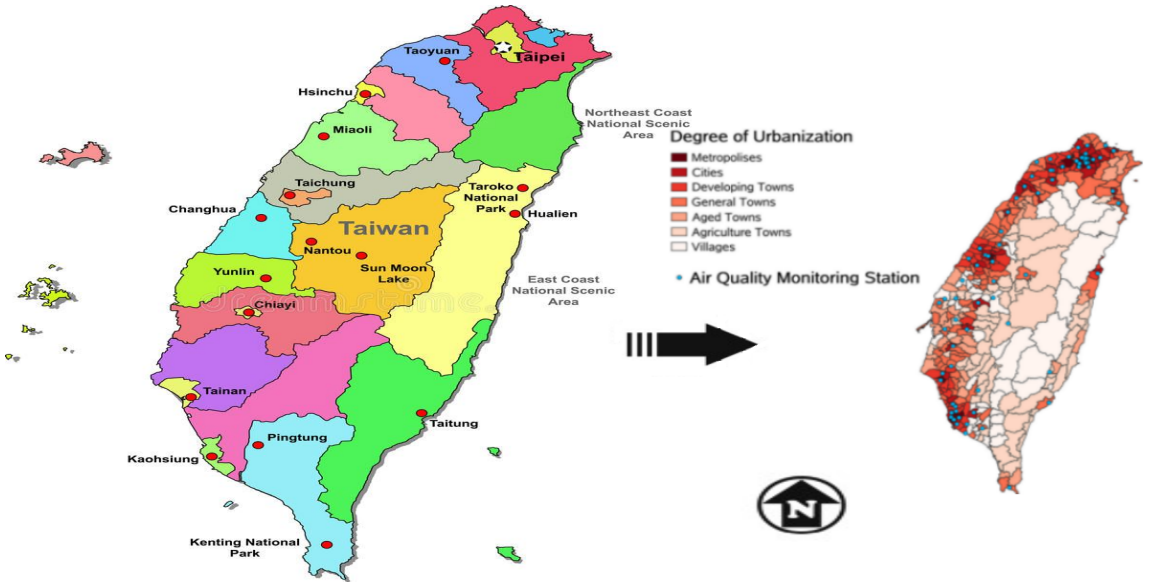

FIGURE 2.3: Location of the data source.

located at latitude $25°4'0''N$ and longitude $121°31'0''E$. Location of the data source as shown in Figure 2.3. The air pollution data available at https://airtw.epa.gov.tw/

2.2.2 Secondary Data

The principle objective of the checking stations is to screen air quality and its drawn-out pattern in a bigger region. We have collected TAQMN data from Taiwan country its consisting different categories of the total 76 air monitoring stations, such as 55 stations are general stations, 6 stations are traffic air quality monitoring stations, 5 stations are industrial air quality monitoring stations, 5 stations are international park observing stations such as H Eng pure and yang name, and last 5 stations are general background stations [27].

2.2.3 Air Pollutants

Air allows our living planet to inhale, it's the combination of gases that fill the air, offering life to the plants and creatures that make Earth a particularly lively spot. Comprehensively talking, the air is essentially comprised of two gases (78 % Nitrogen and 21 % Oxygen), with a couple of different gases (like carbon dioxide and argon) present in totally minute amounts the asset of the air contamination as displayed.

1. **Sulphur Dioxide(SO_2)**:

   Charcoal, diesel, and other non-renewable energy sources are frequently unclean and comprise of sulfur and furthermore normal (graphite-found) parts. Right when sulfur combined with oxygen (O_2) from the air, sulfur dioxide (SO_2) is molded [28].

2. **Carbon Dioxide (CO_2)**:

   This gas is noteworthy to step by step life and isn't ordinarily seen as defilement, we release everything when we breathe in out and plants, for illustration, yields, trees, and boondocks have to be "breathe in" it into creating.

3. **Nitric Oxide (NO)**:

   Nitrogen dioxide (NO_2) and Nitric oxide (NO) are toxins made as an unusual result of burning when nitrogen (N) and oxygen (O_2) from the air react together. Nitric oxide (NO) defilement begins from motors, engines, and power stations and accepts an essential portion within the forming of destructive rain, ozone (O_3), and fog[29].

4. **Particulates Matters ($PM_{2.5}$ and PM_{10})**:

   Particulate Matter ($PM_{2.5}$ and PM_{10}) are the very dangerous particles in air because of its size is very very small, particulate matter ($PM_{2.5}$) its diameter is less than 2.5 microns, and particulate matter (PM_{10}) its diameter is less than 10 microns. It is very harmful for the living organism breathing problems, plants, and destroy the ozone layer within a short period of time. In urban communities most particulates originate from transportation and unstructured industrialization.

2.3 Data Mining Models

This section give a short description of the predictive, forecasting, and clustering for machine learning and deep learning algorithms. Data mining is utilized to portray insight in data sets, it is a strategy of extricating and perceive helpful data and succeeding information from data sets utilizing numerical, measurable, man-made reasoning, and AI procedure. Data mining solidifies numerous different calculations to put through various assignments. This load of calculations acclimatize the

model into the information[30]. The calculations analyze the information and tweak the information that is nearest to the components of the information being inspected.

2.3.1 Linear Regression

Linear regression algorithm is the direct algorithm and its shows the relationship between the target variable and one or more autonomous factors it show the true information to the particular condition[31].

$(x_1, y_1, ..., (x_n, y_n)$

2.3.2 Logistic Regression

Logistic regression may be a quantifiable demonstration that in its principal structure utilizes a key capacity to appear a two-fold dependent variable, yet a part more perplexing developments exist. In regression examination, key regression (or logit regression) is surveying the boundaries of a calculated show (a sort of paired regression).

2.3.3 K Nearest Neighbours (KNN)

The shortened form KNN stands for "K-Nearest Neighbour". It may be a directed machine learning algorithm. The calculation can be utilized to unravel both classification and regression issue statements. The number of closest neighbors to a modern obscure variable that should be anticipated or classified is signified by the symbol 'K' [32].

2.3.4 Random Forest

Random forest may be a gathering of numerous decision trees. Irregular forests are built employing a methodology called stowing in which choice trees are utilized as parallel estimators. In case utilized for a classification issue, the result is based on the larger part vote of the comes around received from each choice tree. For regression, the forecast of a leaf center is the mean regard of the target values in that leaf. Random forest backslide takes the mean regard of the comes approximately from choice trees [33].

2.3.5 Gradient Boosted Decision Trees (GBDT)

GB is a ML procedure for regression, classification, & other tasks, which produces a forecast demonstrate within the frame of an outfit of weak prediction models, regularly choice trees. When a choice tree may be a frail learner, the coming about calculation is called slope boosted trees, which more often than not beats arbitrary forest. It builds the show in a stage-wise mold as other boosting strategies do, and it generalizes them by permitting optimization of a subjective differentiable misfortune work.

2.3.6 Clustering

Clustering or cluster examination could be a machine learning procedure, which bunches the unlabeled dataset. It can be characterized as "A way of gathering the information focuses into distinctive clusters, comprising of comparable information focuses. The objects with the conceivable similarities stay in a bunch that has less or no similarities with another group."

2.4 Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANNs) are quantifiable models explicitly persuaded by and to some degree demonstrated on natural neural frameworks. They can demonstrate and taking care of nonlinear associations among information sources and yields in equal (find in figure **??**). The connected calculations are a piece of the more extensive field of AI and can be used in various applications as examined[34].

2.4.1 Deep Learning

Deep learning (also known as profound coordinated learning) is a part of a more extensive group of AI procedures dependent on counterfeit neural frameworks with portrayal learning. Learning can be coordinated, semi-directed. Deep learning could be a class of AI calculations that utilizes various layers to progressively remove more significant level features from the row information. Most exceptional profound learning models depend on artificial neural networks, especially convolutional

neural networks (CNN)s, notwithstanding the way that they can fuse propositional conditions or inactive elements coordinated layer-wise in significant generative models like the hubs in significant conviction frameworks and significant Boltzmann machines.

2.5 Evaluation Metrics

2.5.1 Performance Metrics of Forecast

The proposed model's tests were done utilizing the dataset chosen from Taiwan country stations. A brief portrayal of the Mean square error (MSE),Mean absolute percentage error (MAPE), Root mean square error (RMSE), Mean absolute error (MAE), and Forecast skill (SS) these measures are utilized to compute the execution of the proposed system through the forecast handle[35]. Let x_i be the i^{th} real value, (xi) be the i^{th} predicted values $\overline{x}$ by the numerical error of observed values and n is the number of observed values. The scientific definition of MSE, MAE, RMSE, MAPE, and SS measures represented by equation.

- **Mean Absolute Error (MAE)**:

  Mean absolute error is the normal of the supreme values of the deviation. This sort of error estimation is valuable when measuring forecast errors within the same unit as the original arrangement equation as 2.1.

$$MAE = \frac{1}{k}\sum_{i=1}^{k}|\varepsilon_t| \tag{2.1}$$

- **Mean Squared Error (MSE)**:

  The MSE is the normal of the square forecast error. As the square of the mistake is taken, the impact of the bigger errors has more weight in the score is 2.2.

$$MSE = \frac{1}{k}\sum_{i=1}^{k}\varepsilon_t^2 \tag{2.2}$$

  This is often one of the foremost measures to assess and discover models, but once the model is found, generally other error measures such as MAE are utilized.

- **Root Mean Squared Error (RMSE)**:

  As RMSE is obvious by the name itself, that it may be a basic square root of mean squared error is 2.3.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\varepsilon_t^2} \tag{2.3}$$

- **Coefficient of Determination** (R^2):

  Frequently utilized within the case of linear regression issues, R^2 determines how much of the full variation in α (dependent variable) is clarified by the variation in X (independent variable) the R^2 equation as fallows 2.4.

$$R^2 = 1 - \frac{\sum(\alpha_a - \alpha_p)^2}{\sum(\alpha_a - \alpha_m)^2} \tag{2.4}$$

  Where α_a is the actual value, α_p is the predicted value, and α_m mean value. The next R^2 is ideal whereas doing straight regression. The extend of R^2 is (- ,1] (don't get confused by the title, R^2, it can be negative as well!). Whereas high R^2 esteem gives you a sense of the goodness of fit of the demonstrate, it shouldn't be utilized as the only metric to choose the finest demonstrate [36]. In case you care about almost the absolute predictions at that point likely it's way better to check RMSE/MAE as well.

2.6 Conclusion

In this chapter, we have presented the dataset to be utilized in this investigation work. A point-by-point portrayal of the Taiwan Air Quality Monitoring Network (TAQMN) dataset is displayed. In addition, we have briefly presented the investigate strategy to be taken after. More particularly, the ML algorithms and DL algorithms utilized in this research work are depicted. The investigate system taken after in this investigation is depicted and the standard execution measurements fundamental to assess the execution of the proposed regression, prediction, and clustering models are moreover highlighted in detail.

Chapter 3

The Multidimensional Data Model to Creating Data Warehouse for Air Pollution

3.1 Introduction

A large volume of data being gathered and stored in a file is a remarkable rate. The administration and handling of an enormous amount of data consume a lot of time and space. Therefore,the procedure to store, maintain, analyze with an easy and useful format is an essential task. We describe a multidimensional data model to store, group and analyze for decision making with meaningful information from historical and current data. Data Warehouse (DW) is defined as "an accumulation of theme situated, coordinated, unstable, with instance differing information to help the decision procedures of an association". Data warehouse regularly coordinates and stores huge data, indexes from various and heterogeneous data sources and data warehouse provides and promotes accessing schema for fetching time variant data at users convenience.

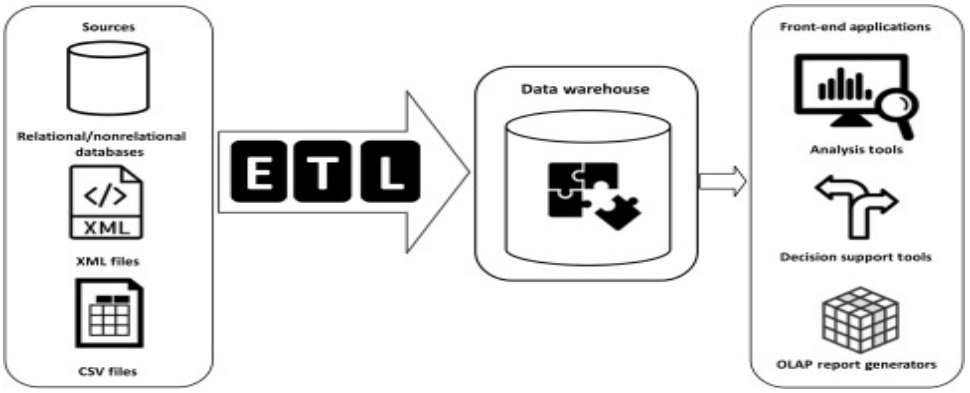

FIGURE 3.1: The basic structure of data warehouse.

3.2 The Proposed Method

Data Warehouse (DW) is defined as "an accumulation of theme situated, coordinated, unstable, with instance-differing information to help the decision procedures of an association". Data warehouse regularly coordinates and stores huge data, indexes from various and heterogeneous data sources and data warehouse provides and promotes accessing schema for fetching time variant data at users convenience [37].

Data warehouse likewise sorts out information as per the fundamental examination subjects of the association in a way that encourages information investigation errands the basic structure of data warehouse as shown in figure 3.1. For instance, a data warehouse of air pollution brought to contain information about investigation subjects, measure units, hours, stations, and so on. This schema define two fundamental information structures, fact tables that store keys for dimensional tables and measures of the examination subjects and the dimension tables that store information identified with investigation criteria [38]. The components of the data warehouse and its sub components are described for facilitating the design of online analytical processing operations on the data warehouse schema.

3.2.1 Multidimensional Data Model

A multidimensional data model is a technique it is used as a database for arranging the data elements in a specified format to show the data in an n-dimensional way.

A multidimensional data model as receiving different queries from end-users, that queries related solutions providing to end-users for particular queries. All queries-related solutions are providing in a short period of time without ambiguity and it also reduces the complexity of the problems. The multidimensional data model was used to implement the OLAP and data warehouse. It is also used to show the data in n-dimensional visualization and easy to elaborate the data in different ways [39].

- **Meta Data**

  Meta data is characterized as the information, giving data around one or more angles of the information. It is utilized to summarize fundamental data approximately.

- **Data Cube**

  A data cube is characterized by measurements and realities. The information, the data cube is utilized and speak to information along a few degrees of intrigued. The cells within the data cube and their computed aggregated values are visualized and shown in the figure 3.5.

- **Data Mart**

  It is centered on single useful range of a information and contains a subset of information put away in a data warehouse. Information Shop makes a difference to upgrade user's reaction time due to admonishment in the volume of information.

3.2.2 Steps in Multidimensional Data Model Analysis

Data warehouse and data cube depend on the multidimensional data model. This model considers decisions information (key, data) as per the perspective of the chiefs by sorting out it multidimensional into facts and dimensions. First speaks to subjects of examination and is depicted by properties (by and large numerical) called measures. For instance, the model describes in the realities "Air pollution" are portrayed by temperature (T), carbon monoxide (CO), particulate matter (PM_{10}), sulfur dioxide (SO_2) and wind speed (WS) measures. Dimensions speak to gauge investigation criteria and take into consideration seeing measures from alternate points of view. In the case of air pollution measures are broken down as per time and station dimensions.

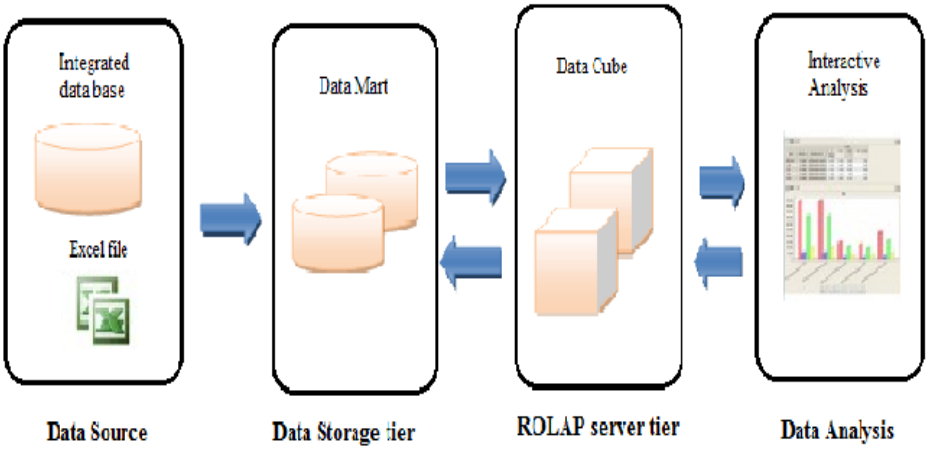

FIGURE 3.2: Multidimensional Data Model.

As a data source, it containing the combined data and work sheets and extraction in the data storage tier it as transform into the server tier, then build a data warehouse and cube structure then analyzing information to produce the interactive analysis shown in the Figure 3.2.

- **Schema**

  The data warehouse build by multidimensional data model is a coherent plan system at looks to show information in an instinctive shape and with a superior. Single multidimensional ideal is made out to record by different remote key, called reality table, and an arrangement of literal records is a dimension tables.

- **Star Schema**

  The outline of the star schema it contains the major table as air pollution table and smaller tables are namely stationed, pollutants, and time table. In the air pollution table it consists of remote keys such as pollutant key, station key and time key and also it contains the measures such as average, minimum, maximum, standard deviation, and coefficients of variations and above keys are act as a foreign key to the remaining tables [40]. 3.3.

- **Snowflake Schema**

  The snowflake schema is a logical arrangement of fact tables and dimensional tables in the air pollution multidimensional data model. In a star schema also containing the one fact table and remaining dimensional tables, but in this schema dimensional tables can be normalized into some other dimension tables it can be easy to access the data and size of the memory

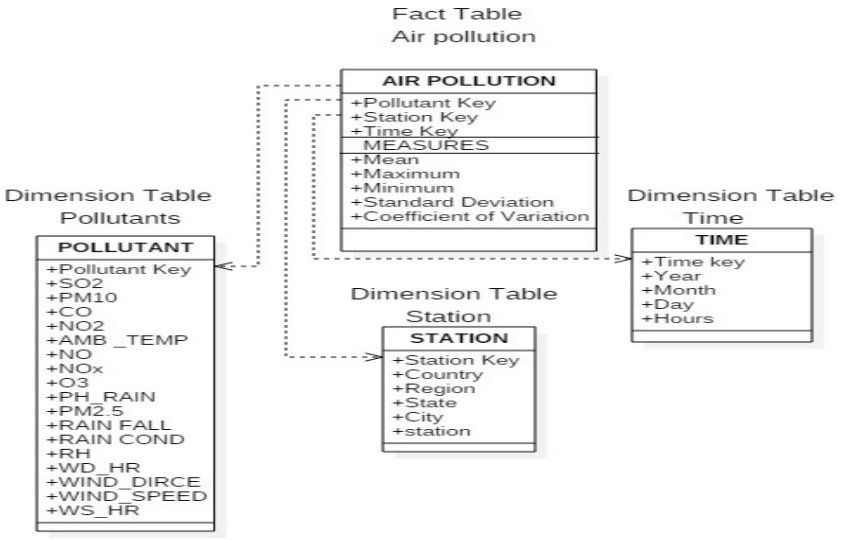

FIGURE 3.3: Star Schema for air pollution.

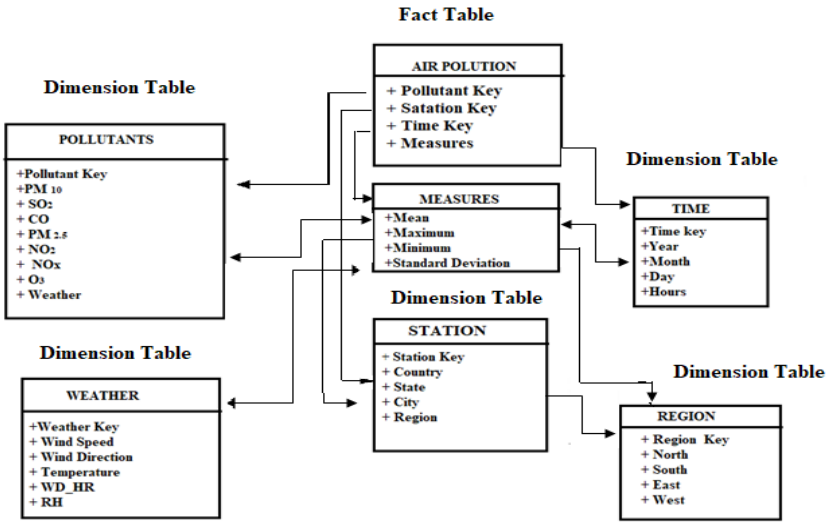

FIGURE 3.4: Snowflake schema for air pollution.

can be decreased in a centralized fact table containing a primary keys are like pollutant key, station key and time key and also it containing the some measures such as average, minimum, maximum, standard deviation, and coefficients of variations.

- **Fact Table**

  Fact tables record estimations or measurements for a particular occasion. Actuality tables, for the most part, comprise of number qualities, the foreign key (Fk) to measurement information where clear data are reserved. literal records are intended less amount of information, which means certainties can record occasions at an extremely small level. This can bring about

the amazing of countless in a reality table after some time. In the air pollution fact table, it consists of pollutant (Fk), station (Fk), time (Fk) and its also included some measures such as aggregation, maximum, minimum, and standard deviation [41].

- **Dimension Table**

  As dimension tables handles elaborate variable data, the number of samples inputted are comparatively more than the fact table. The dimensions in a multidimensional data model of a data warehouse clarifies a different qualities. Some of the repeatedly used dimension tables are time table, pollutant table and station table. Moreover, the tables are also assigned a primary key referring to the grouping of attributes to create the primary key. Example, to create an air pollution table referring to an event with pollutant key, station key, time key, pollution data generated.

3.3 Experimental Results

Air pollution data was collected from Taiwan country in the year from 2012 to 2017 its total of 5 years of data in this data is time-series data because every day we collected data. It includes air pollution particles and meteorological data and data is generated every one hour through particular sensors.

Data warehouse is the most important things because of its made-up off data mart, metadata, and data cube. Air pollution data can be represented in the form of data cube using descriptive statistical analysis, these statistical analysis by using this equations the mathematical definition of mean (μ), standard deviation (σ), and coefficients of variations (CV) measure represented by the Equations (3.1), (3.2), and (3.3) respectively.

$$\mu = \sum_{i=1}^{n} \frac{x_i}{n} \tag{3.1}$$

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2} \tag{3.2}$$

TABLE 3.1: Statistical analysis values determined the level of air pollutant emission in Taiwan country.

Measures	Mean	Min	Max	Stdiv	CV
AMB_TEMP	15.43844	10.00	26.0	2.1583	13.980039
CO	0.46043	0.18	1.1	0.1299	28.212758
NO	2.18440	0.10	33.0	2.2282	102.005127
NO2	15.20200	6.20	47.0	4.7601	31.312327
NOx	17.39580	7.00	61.0	6.0604	34.838294
O3	23.73683	1.20	74.0	12.5652	52.935459
PH_RAIN	0.30080	0.00	2.7	0.3855	128.158245
PM10	78.05100	16.00	215.0	23.8082	30.503389
PM2.5	46.50940	7.00	131.0	13.9585	30.012213
RAINFALL	0.00800	0.00	2.7	0.0252	315.000000
RAIN_COND	0.93360	0.00	2.7	2.8884	309.383033
RH	69.01880	28.00	95.0	8.8826	12.869827
SO2	3.69270	0.50	13.0	1.5485	41.934086
WD_HR	248.99010	0.30	358.0	76.2235	30.613065
WIND_DIREC	242.44030	2.10	358.0	83.1889	34.313148
WIND_SPEED	2.93760	0.50	8.3	0.8907	30.320670
WS_HR	1.35380	0.10	3.9	0.7488	55.310977

$$CV = \frac{\sigma}{\mu} * 100 \tag{3.3}$$

The statistical analysis of five years of data from 2012 to 2017 for Taiwan country all stations are statistical evaluation such as minimum, maximum, mean, standard deviation, and coefficient variation the descriptive statistics values determining the level of pollutant emissions as shown in Table 3.1. Further, we have to determine each and every stations statistical values calculated separately such as Annan, Chiayi, Good, Mumng, and Newport stations. The data warehouse is a collection of data cube, So all stations collected data, with calculated statistical values are show data cube for air pollution pollutants, stations, and time period for year as shown in Figure 3.5.

The statistical analysis for minimum, maximum, mean, standard deviation, and coefficient variation calculation for all stations data are shown in line Figure 3.6.

3.4 Conclusion

In this chapter we developing the Taiwan country air pollution data warehouse, the multidimensional data model is used to implementation of the data warehouse its collection of data cube, data

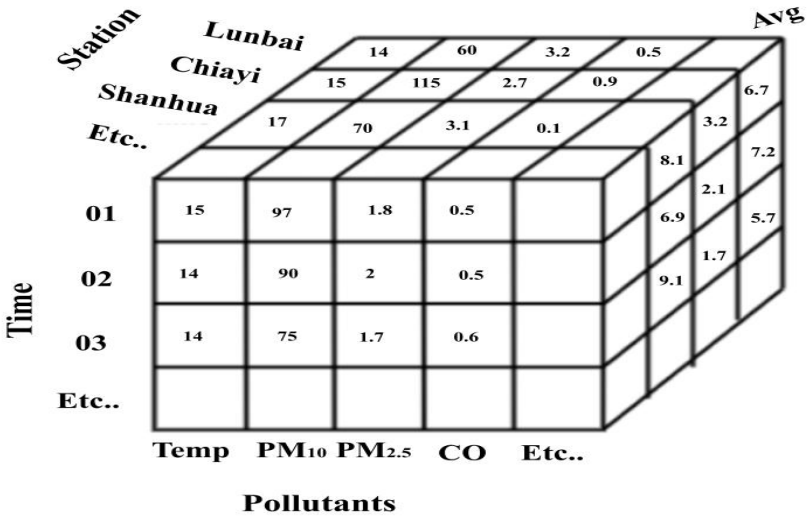

FIGURE 3.5: Data cube visualization for air pollution.

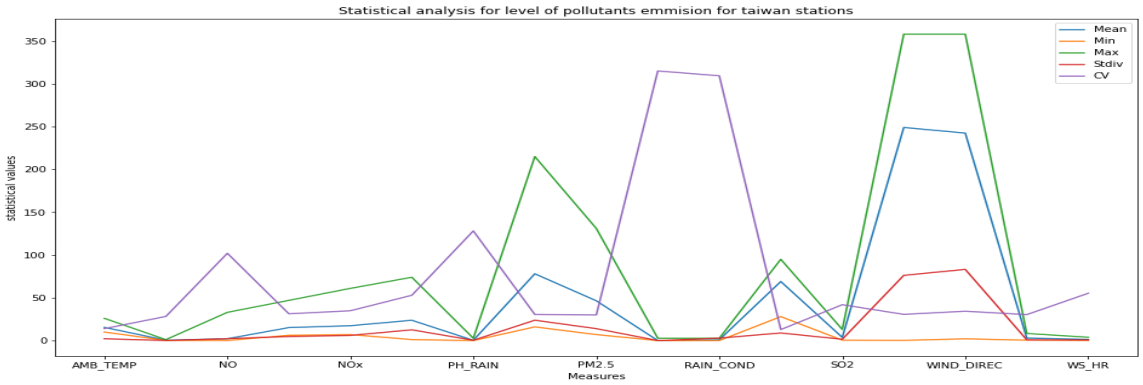

FIGURE 3.6: Statistical analysis values determined by line graph.

mart, and meta data using star schema and snowflake schema the statistical analysis of five years of data of Taiwan country all stations are statistical evaluation such as minimum, maximum, mean, standard deviation, and coefficient variation the descriptive statistics values determining the level of pollutant emissions. In selected stations are considered for resolving decision making issues, it can be easy to access the data and size of the memory can be decreased.

Chapter 4

Visualization of Framework and Correlation Analysis for Air Pollution

4.1 Introduction

Taiwan country has faced an extreme fine particulate matter ($PM_{2.5}$) issue. Until this point, impressive advancement understanding the $PM_{2.5}$ issue, counting spatial-worldly portrayal, driving elements, and well-being impacts. Be that as it may, little exploration has been done on the unique collaborations and connections between $PM_{2.5}$ fixations in various urban communities around here. To address the investigation gap, the analysis of time-slacked intercity correlations of $PM_{2.5}$ time series and proposed a visualization system dependent on this peculiarity to imagine the collaboration in $PM_{2.5}$ focus between urban areas [42].

The representations additionally show that the relationships are more critical in winter seasons and between urban areas that are nearer and occasional changes in the worldly request of the connected $PM_{2.5}$ time series. Assist analysis proposes that the time-slacked intercity relationships of $PM_{2.5}$ time series are undoubted because of succinct meteorological varieties [43]. In this work, we purposed a visualization technique based on this principle of the visualization, cross-correlation method, and also the time-series concentration with particulate matter $PM_{2.5}$ for different cities in Taiwan. The visualization also shows that the correlation between the different meteorological factors as well as the different air pollution pollutants for particular cities in Taiwan. This visualization

approach helps to determine the concentration of the air pollution levels in different cities and also determine the Pearson correlation r values for selected cities.

4.2 Visualization Framework

The process of configuration can be isolated into four sections. The initial step was the information preprocessing, as referenced previously. Then, at that point, directing a basic examination of the spatio-temporal appropriation example of $PM_{2.5}$ concentration. Thirdly, directing the connection examination at various scales, i.e., city, district, season, and year. The connections between day by day $PM_{2.5}$ fixation and every day meteorological variables were estimated by Spearman's position relationship coefficient. At last, we examined the occasional and territorial varieties of the connections between $PM_{2.5}$ concentration and the meteorological elements. Figure 4.1 shows the proposed model.

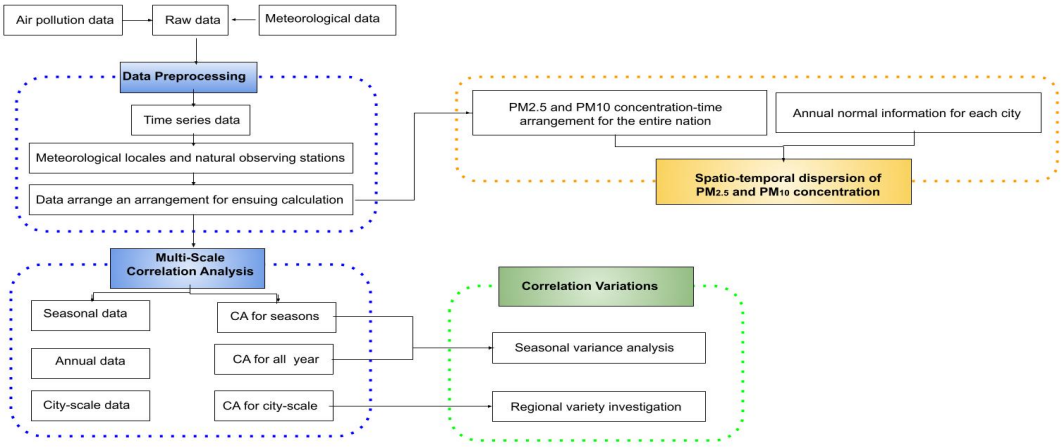

FIGURE 4.1: Architecture of the the proposed model

4.2.1 Correlation coefficients

Correlation coefficients are utilized to quantify how solid a relationship is between two factors [44]. There are a few sorts of connection coefficients, yet the most well-known is Pearson's. Pearson's relationship (likewise called Pearson's r) is a connection coefficient generally utilized in direct relapse.

Pearson's relationship coefficient equation is the foremost commonly used. In case it takes a fundamental details class, this is one of the presumably utilized equation 4.1.

$$r = \frac{n\sum \alpha_i\beta_i - (\sum \alpha_i)(\sum \beta_i)}{\sqrt{\left[n\sum \alpha_i^2 - (\sum \alpha_i)^2\right]\left[n\sum \beta_i^2 - (\sum \beta_i)^2\right]}} \tag{4.1}$$

4.3 Experimental Results

The collected data from January 2012 to December 2017 that were obtained by Taiwan air quality monitoring network(TAQMN). It as hourly data set it can be converting into a daily, monthly, and yearly data set.

This method visualization framework to visualize the continue interactions of particulate matter ($PM_{2.5}$) as well as all other pollutants of air pollution and meteorological factors time series in different cities by the reference of TAQMN data as shown in figure 4.2 and figure 4.3.

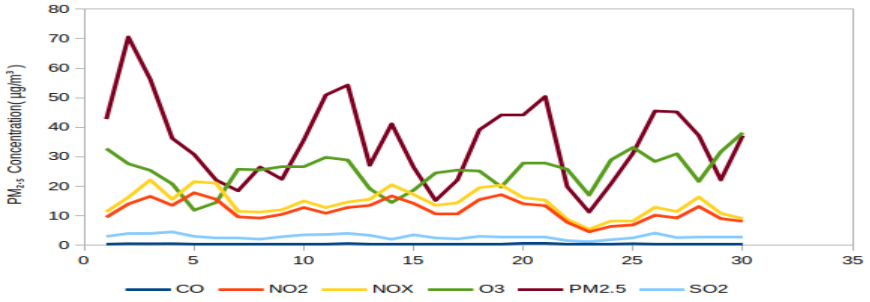

FIGURE 4.2: The time-legged correlation of ($PM_{2.5}$) with air pollutants

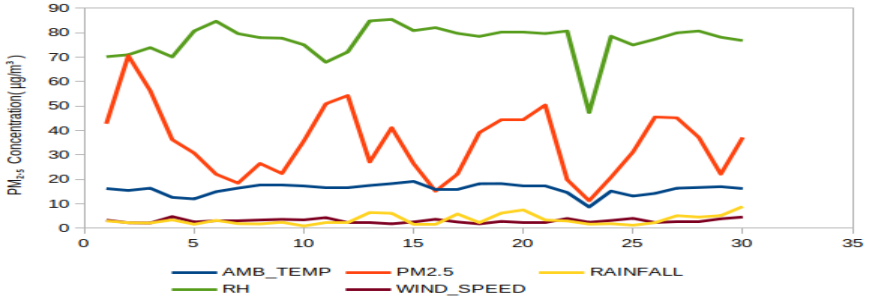

FIGURE 4.3: The time-legged correlation of ($PM_{2.5}$) with meteorological factors

Similarly the visualization framework for particulate matter (PM_{10}) as shown in figure 4.4 and figure 4.5. The important things in the visualization framework are the cross-correlation method [45], it calculates the correlations between the particulate matter ($PM_{2.5}$) and particulate matter

(PM_{10}), air pollutants and meteorological factors time series in different cities as shown in figure 4.6 and figure 4.7 respectively.

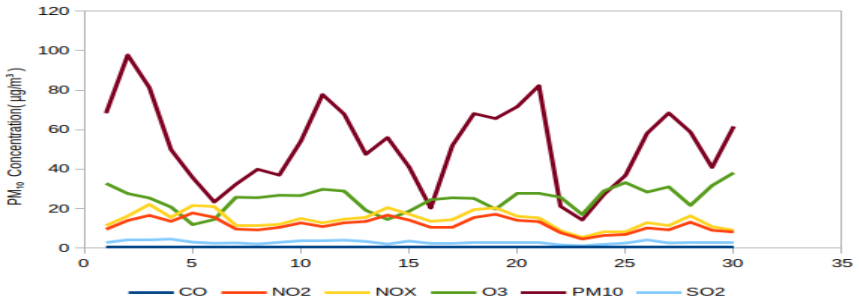

FIGURE 4.4: The time-legged correlation of (PM_{10}) with air pollutants

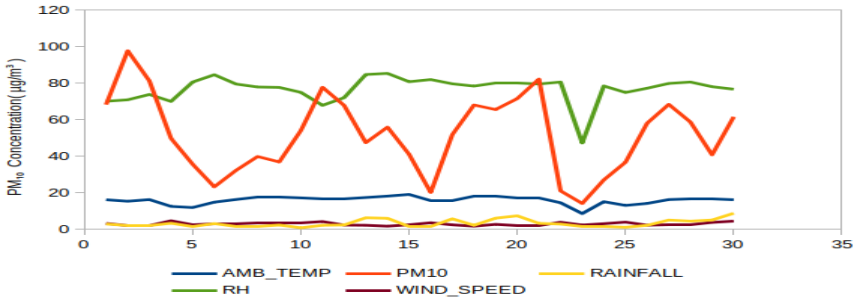

FIGURE 4.5: The time-legged correlation of(PM_{10}) with meteorological factors

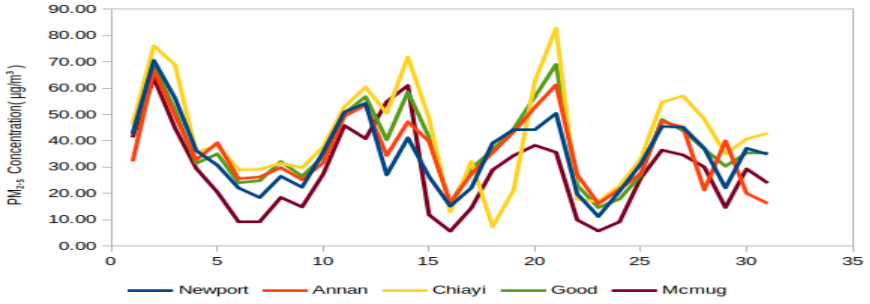

FIGURE 4.6: Particulate Matter ($PM_{2.5}$) concentration level analysis for the different stations

The visualization framework is the comparison of time-legged correlation method, it calculates the correlations between the particulate matter ($PM_{2.5}$), air pollutants and meteorological factors time series in individually different pollutants particles such as O_3, NO_x, SO_2, NO_2 with particulate matter ($PM_{2.5}$) and meteorological factors such as rainfall, temperature, wind speed, and relative humidity with particulate matter ($PM_{2.5}$) as shown in figure 4.8 and figure 4.9 respectively.

Pearson correlation (r) is utilized to quantify the strength and bearing of a direct connection between two factors. Numerically this should be possible by isolating the covariance of the two factors by the result of their standard deviations [46]. The worth of r runs between - 1 and +1. A relationship of

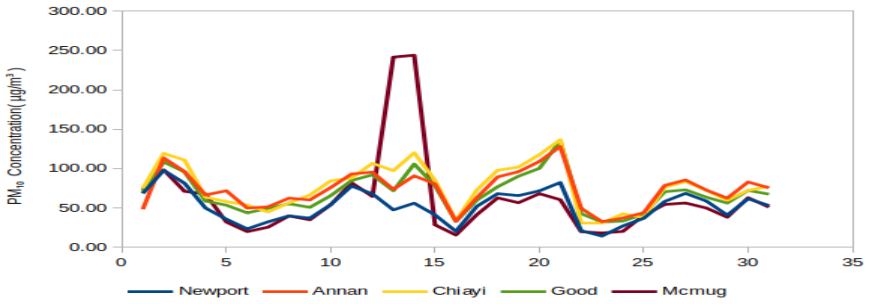

FIGURE 4.7: Particulate Matter (PM_{10}) concentration level analysis for the different stations

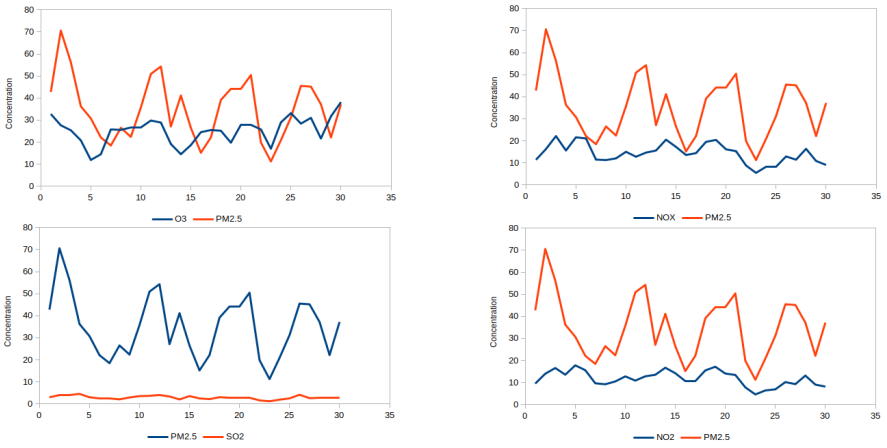

FIGURE 4.8: The comparison of time-legged correlation of ($PM_{2.5}$) with air pollutants

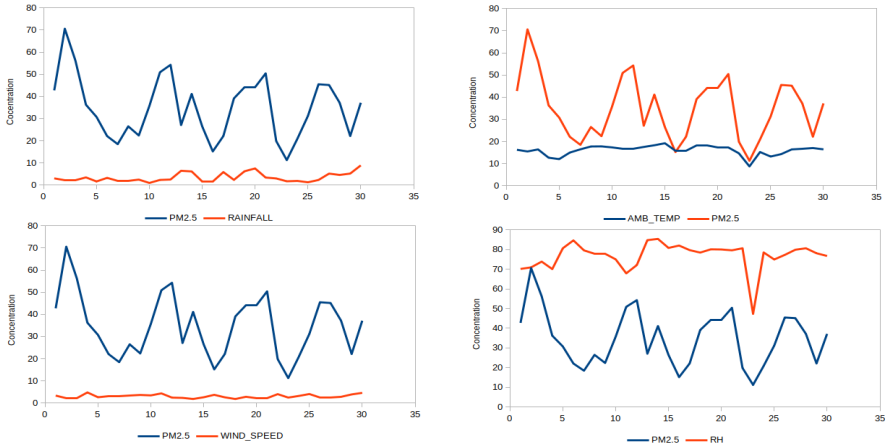

FIGURE 4.9: The comparison of time-legged correlation of ($PM_{2.5}$) with meteorological factors

- 1 shows an ideal negative correlation, while a connection of +1 shows an ideal positive correlation. A connection of 0 shows no connection between the development of the two factors.

Utilize Pearson's correlation coefficient to analyze the strength and course of the monotonic connection between two nonstop or ordinal factors particulate matter and wind speed [47]. In a monotonic

TABLE 4.1: Pearson's r values correlation for different cities

Sl No	Cities	Pearson's Value		Correlation	Strength of Association
		r	p		
1	Annan	0.29	2.1e-42	Positive	Small
2	Mumcg	-0.13	1.5e-09	Negative	Small
3	Good	-0.53	8.4e-156	Negative	Large
4	Newport	0.26	3.7e-36	Positive	Small

relationship, the factors will generally move a similar relative way. To compute the Pearson's correlation as displayed in table 4.1. The Pearson's Correlation for particulate matter($PM_{2.5}$) with wind speed to the various stations like Annan, Mumcg, Good, Newport, in the Annan station correlation as positive and strength of the relationship between particulate matter ($PM_{2.5}$) with wind speed is little, comparatively Newport station additionally positive correlation and strength of affiliation are likewise more modest than staying two stations, for example, Mumcg and Good stations are Pearson's relationship are negative correlation and strength of the relationship between particulate matter ($PM_{2.5}$) with wind speed is little and long with individually [48]. Pearson's Correlation for particulate matter($PM_{2.5}$) with wind speed to the various stations as displayed in figure 4.10.

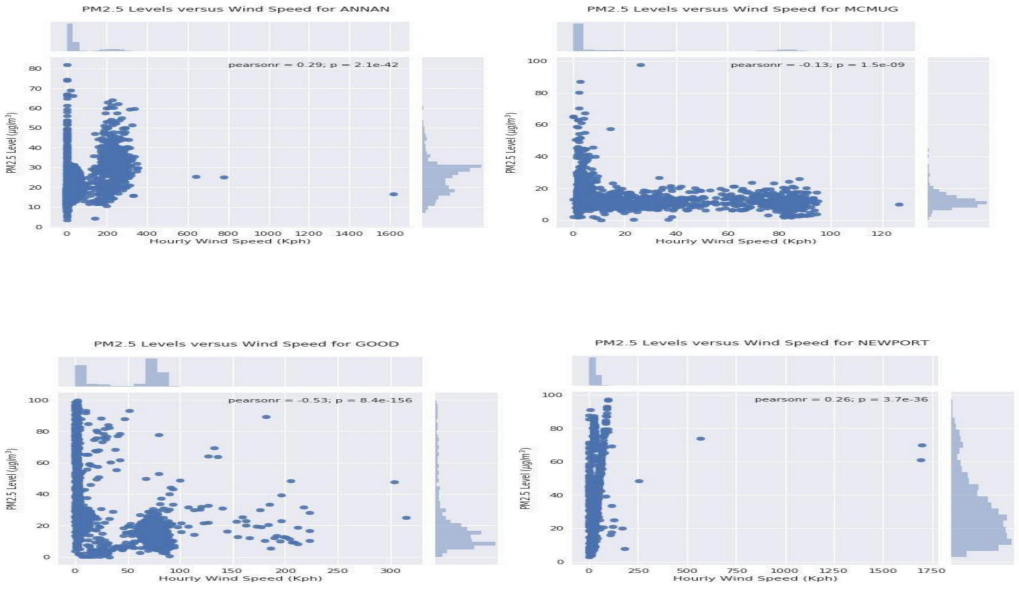
PM2.5 Levels versus Wind Speed for ANNAN
pearsonr = 0.29; p = 2.1e-42
PM2.5 Level (µg/m³)
Hourly Wind Speed (Kph)
PM2.5 Levels versus Wind Speed for MCMUG
pearsonr = -0.13; p = 1.5e-09
PM2.5 Level (µg/m³)
Hourly Wind Speed (Kph)
PM2.5 Levels versus Wind Speed for GOOD
pearsonr = -0.53; p = 8.4e-156
PM2.5 Level (µg/m³)
Hourly Wind Speed (Kph)
PM2.5 Levels versus Wind Speed for NEWPORT
pearsonr = 0.26; p = 3.7e-36
PM2.5 Level (µg/m³)
Hourly Wind Speed (Kph)

FIGURE 4.10: Pearson's Correlation for ($PM_{2.5}$) with wind speed to the different stations

Figure 4.11 shows the visualization framework for the correlation matrix for Annan city in this matrix all show the air pollutants and meteorological factors in this city as wind speed and wind speed horizontal as positive correlation and temperature, carbon monoxide components are the negative correlation. Another station Good city temperature, carbon monoxide, nitrogen dioxide (NO_2),

and wind direction are the negative correlation, and other pollutants and meteorological factors are a positive correlation. Third McMug city correlation matrix ozone (O_3) and particulate matter ($PM_{2.5}$, PM_{10}) are the positive correlation and other pollutants and meteorological factors are a negative correlation. Last city as Newport the correlation matrix nitrogen dioxide (NO_2), sulfur dioxide (SO_2) and wind speed is the positive correlation and other pollutants and meteorological factors are the negative correlation.

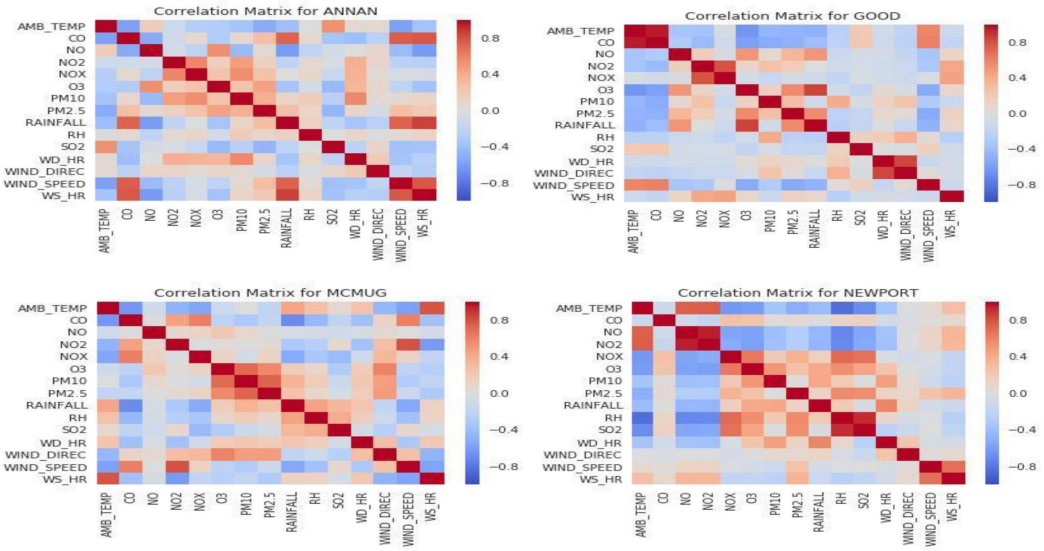

FIGURE 4.11: Correlation Matrix for all air pollutants with different stations

4.4 Conclusion

In this chapter, we outlined the concept of investigating the phenomenon of time series of correlation analysis of the particulate matter ($PM_{2.5}$) with selected cities by using the visualization framework with the cross-correlation analysis method we analyzed the Annan, Newport, Good, and McMug cities are positive and a negative correlation of the air pollution pollutants and meteorological factors. Some cities are showing the maximum positive correlation and less a negative correlation. Visualization framework and a cross-correlation method determine the particulate matter ($PM_{2.5}$) and wind speed for selected cities Annan, Newport, Good, and McMug first two cities are a positive correlation and the other two cities are a negative correlation, then Pearson's Correlation r and p for particulate matter ($PM_{2.5}$) with wind speed to the different stations. In the future we analysis the time series analysis for prediction using machine learning models.

Chapter 5

Time Series Analysis for Prediction of Particulate Matter ($PM_{2.5}$) Using SARIMA Model

5.1 Introduction

Air pollution is alarmingly increasing in many parts of the world. This is due to several reasons such as industrialization and population growth. At the end of this decade, the number of people becomes seven billion globally hence, the number of people living in the urban areas is also gradually increasing. Particulate matter (PM) is the most important pollutants stirring of the air quality in Taiwan and the rest of the world which is a complicated particle of the liquid and solid components [49]. Moreover, biological, human activities and chemical substances in the air are very harmful components for human health and all living animals on earth. ($PM_{2.5}$) arises from many sources such as harmful activities human belongs, industries, automobiles, urbanization's, power plants, chemical activities and some of the other natural activities such as agricultural burning, and volcanic eruptions [50].

5.2 Time Series Models

The core concept of the time series modeling towards forecasting trends in that will occur in the future from data, and the most popular models used for this purpose is the box-Jenkins ARIMA Model [51]. ARIMA is a very popular statistical method for time series forecasting. The basic idea of ARIMA model is based on the following assumptions.

1. The data series under consideration is stationary, in a sense that the mean and variance does not vary with time.
2. The input data must be a uni variate series as SARIMA uses past values to predict future values.

5.2.1 Seasonal ARIMA (SARIMA (p, d, q), (P, D, Q))

Regular time arrangement information can be examine via Box-Jenkins technique. When a model comprises set of standard components, then the model is Seasonal Autoregressive Integrated Moving Average (SARIMA). The regular ARIMA model is expressed as SARIMA (p, d, q) (P, D, Q)$_S$, where the lower cases appears to be nonseasonal portion and the capitalized letters become seasonal portion. SARIMA (p, d, q), (P, D, Q)$_S$ can be composed as equation 5.1.

$$\varphi_p(B)\varphi_p(B^s)(1-B)^d(1-B^s)^D Y_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t \tag{5.1}$$

5.3 Experimental Results

In this part, tests for characterizing TAQMN datasets were exhibited by applying the time series model that worked on the exactness prediction of air pollution data. Exploratory consequences of the proposed model are examined in the accompanying subsections.

We collected the data in the form of hourly then it's converted into daily and monthly basis. The steps involved to select the best model is presented in Figure 5.1. Moreover, Figure 5.2 presents the monthly average of $PM_{2.5}$ from January 2012 to December 2017, that shows the different variations for different years.

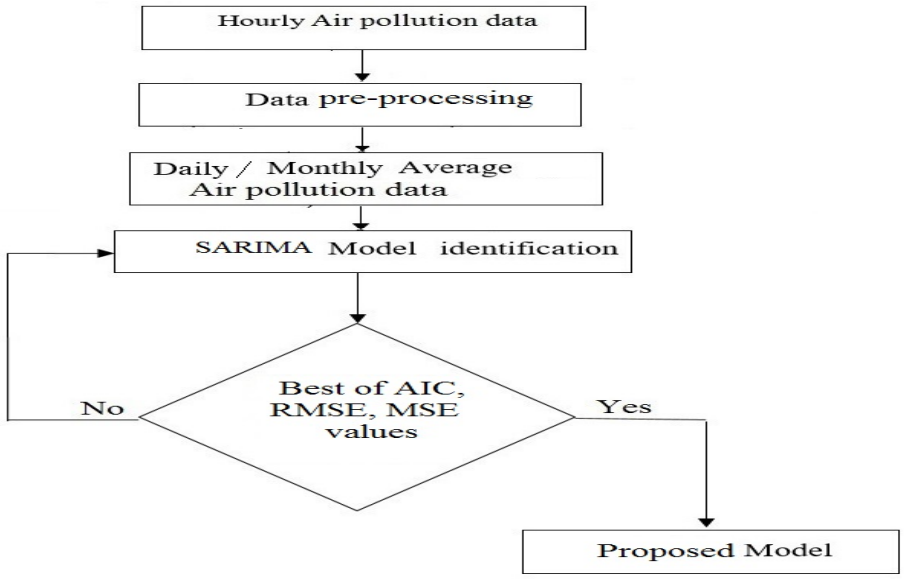

FIGURE 5.1: A flowchart for the proposed procedure

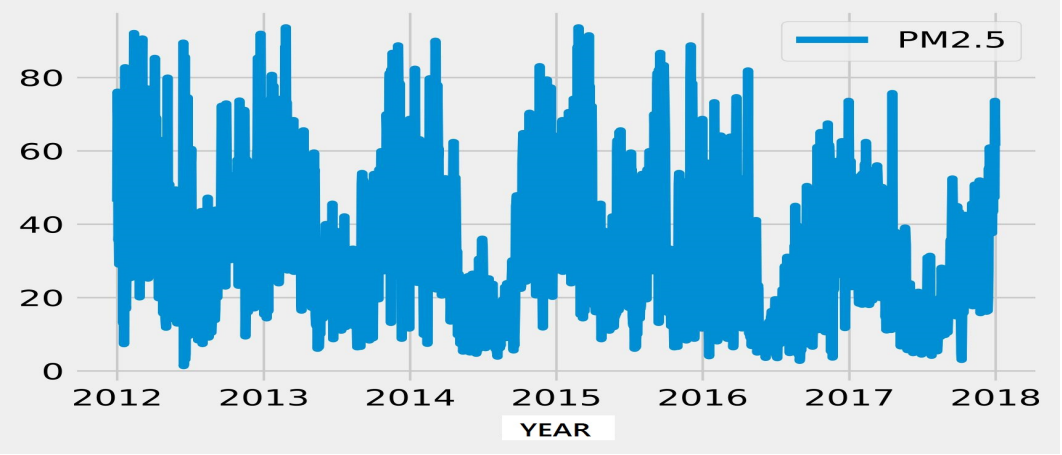

FIGURE 5.2: A time series plot of the monthly average of $PM_{2.5}$ in ($\mu g/m^3$) of TAQMN data period of 2012 to 2017.

5.3.1 Evaluation Functions

The primary criterion used for selecting the best ARIMA model is Akaike Information Criteria (AIC) [52]. The AIC is measure of the relative quality of different statistical methods and among the numerous values of AIC, the least value of AIC is associated with the proposed model which is used for forecasting[53]. Some of the statistical evaluation methods are used to evaluate the model's performance such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Percentage Error (MPE), Mean Square Error (MSE), and Root Mean Square Error (RMSE). The Akaike's Information Criterion (AIC) can be represented mathematically as shown in equation 5.2.

$$AIC = -2\log(L) + 2h \tag{5.2}$$

Where, h is the number of parameters and L is the log- likelihood within the show i.e., h = (p+d+q+P+D+Q). The criteria formulas are shown in equations (5.3) to (5.8).

$$MAE = \frac{\sum_{t=1}^{n} |y_t - \hat{y}_t|}{n} \tag{5.3}$$

$$MAPE = \frac{\sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right|}{n} \tag{5.4}$$

$$MPE = \frac{100\%}{n} \sum_{t=1}^{n} \frac{y_t - \hat{y}_t}{y_t} \tag{5.5}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_t - \hat{y}_t)^2 \tag{5.6}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{n}} \tag{5.7}$$

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{5.8}$$

Where y_t, $\hat{y}_t$ and n stands for the actual, predicted values and number of samples respectively. Moreover, the r_{xy}, and corresponds to the correlation between the x and y variables, average values of the x and y respectively. Finally the x_i and y_i stands for the i^{th} values of either variables respectively.

5.3.2 Experiments

The TAQMN data set is divided into training data set from the year 2012 to 2016 and testing data set from 2016–2017. The highest $PM_{2.5}$ concentration of air pollution was 161 $\mu g/m^3$ in a Chiayi station particular place from the year 2012 to 2017. The following parameter values are used to select the best ARIMA model: the range of p parameter is from 0 to 2, q from 0 to 2 and difference d as 1 and P its values ranges from 0 to 2, Q its ranges from 0 to 2, D its also value as 1, and S is the length of seasonal period 12. Table 5.1 shows the top three AIC, RMSE, and MSE values for the different SARIMA models.

TABLE 5.1: The Best AIC Values For SARIMA Model

Model : SARIMA			
Parameters[(p,q,d),(P,D,Q)]	**MSE**	**RMSE**	**AIC**
[(0, 1, 2), (2, 0, 0, 12)]	0.2989	0.5467	43.767
[(0, 1, 2), (2, 0, 1, 12)]	0.2968	0.5448	41.589
[(2, 1, 2), (1, 0, 0, 12)]	0.2986	0.5464	48.514

TABLE 5.2: The best result of diagnostics test for SARIIMA models for the p-values less then 0.05

	(p,d,q)	**(P,D,Q)**	**AIC**	**MAPE**	**ME**	**MAE**	**MPE**	**MSE**	**RMSE**	**CORR**
6	(0,0,3)	(0,1,1,12)	33.42	0.05	0.14	0.15	0.05	0.04	0.20	0.96
0	(0,0,1)	(0,1,1,12)	38.68	0.07	0.21	0.21	0.07	0.07	0.26	0.94
13	(0,1,2)	(0,0,1,12)	53.55	0.12	0.13	0.34	0.06	0.19	0.44	0.63
17	(0,1,3)	(0,0,1,12)	57.37	0.14	0.29	0.39	0.11	0.26	0.51	0.58
3	(0,0,2)	(0,1,0,12)	60.92	0.02	-0.001	0.07	0.0006	0.01	0.10	0.97
11	(0,1,0)	(0,0,1,12)	60.97	0.16	0.44	0.46	0.16	0.37	0.61	0.54

The AIC values for different SARIMA models have obtained automatically. In Figure 5.3 show the The ACF and PACF monthly average of $PM_{2.5}$ in () of TAQMN data from of 2012 to 2017, and Table 5.1 shows all SARIMA models that have p values less than 0.05. From those values the finest SARIMA demonstrate is chosen based on the taking after measures such as MAPE, MAE, MPE, MSE, RMSE, and relationship coefficient values which are summarized in Table 5.2.

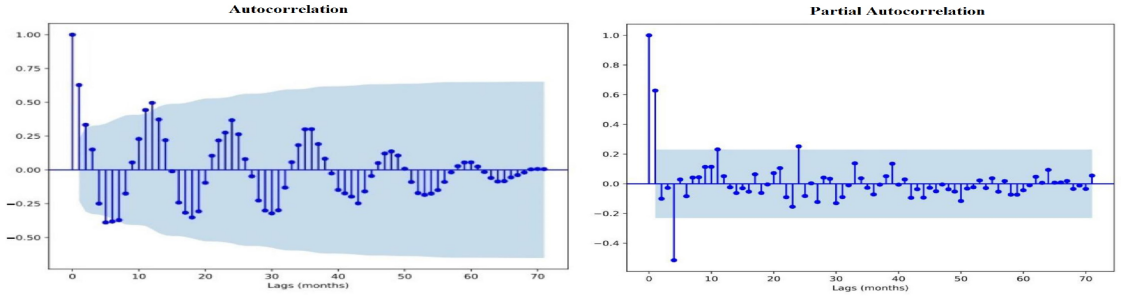

FIGURE 5.3: The ACF and PACF monthly average of $PM_{2.5}$ in() of TAQMN data from of 2012 to 2017

The proposed SARIMA (0, 0, 2) × (0, 1, 0, 12) model are trained by the TAQMN training data as well as testing data set to obtain good result for predicted values as shown in Table 5.3. Figure 5.4 presented the actual and predicted values for the training data set with MSE 2.75 and for testing MSE is 83.53. Moreover, the blue color dotted line indicates the actual values and a red dotted line indicates the prediction of the particulate matter $PM_{2.5}$ of TAQMN data.

TABLE 5.3: The result of prediction of the SARIIMA model (0, 0, 2) x (0, 1, 0, 12) it including the lower and upper values of particulate matter $PM_{2.5}$

Date	Actual	Lower	Predict	Upper
31-01-2016	111.3295	80.8727	99.2642	117.6558
29-02-2016	121.7375	86.3354	104.7267	123.1179
31-03-2016	119.9879	89.3531	107.7441	126.1351
30-04-2016	110.1784	105.2518	123.6427	142.0335
31-05-2016	100.8551	110.8503	129.2411	147.6318
30-06-2016	86.8623	104.273	122.6637	141.0544
31-07-2016	93.9645	109.0798	127.4704	145.8611
31-08-2016	104.0555	101.1023	119.4929	137.8835
30-09-2016	106.2734	111.2583	129.6489	148.0395
31-10-2016	113.2115	98.5274	116.918	135.3086
30-11-2016	118.9314	74.9061	93.2967	111.6873
31-12-2016	124.3604	84.4782	102.8688	121.2594
31-01-2017	119.3152	91.766	110.1566	128.5472
28-02-2017	121.9759	69.4995	87.8901	106.2807
31-03-2017	120.816	110.7343	129.1249	147.5155
30-04-2017	112.7527	99.511	117.9016	136.2921
31-05-2017	104.568	94.7743	113.1649	131.5555
30-06-2017	87.7721	111.462	129.8526	148.2431
31-07-2017	95.5339	104.5489	122.9394	141.33
31-08-2017	97.9445	102.3118	120.7024	139.0929
30-09-2017	111.633	110.7213	129.1119	147.5025
31-10-2017	110.5878	88.058	106.4486	124.8392
30-11-2017	114.0331	89.1953	107.5859	125.9765
31-12-2017	121.4048	97.8474	116.238	134.6286

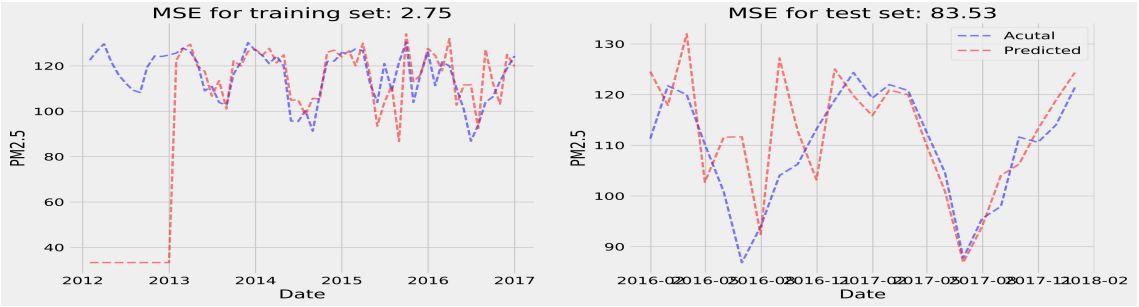

FIGURE 5.4: Plots of actual and predicted MSE for the training set from 2012 to 2016 and MSE for testing set 2016 to 2017.

Figure 5.5 (a) shows standardized residuals over time for seasonality with white noise, (b) correlogram for particulate matter $PM_{2.5}$ histogram plus estimated curve as red colour kernal density estimation (KDE) slightly interface with the N(0,1) green curve, both curves and histogram as height interface in zero point. In (c) it can be represented the normal Q-Q plot, x axis as theoretical quantiles and y axis sample quantiles follows linear trend except small number of points, (d) auto correlation it as correlation between the residual points and lagged points its ranges shows in

between +0.5 to -0.5.

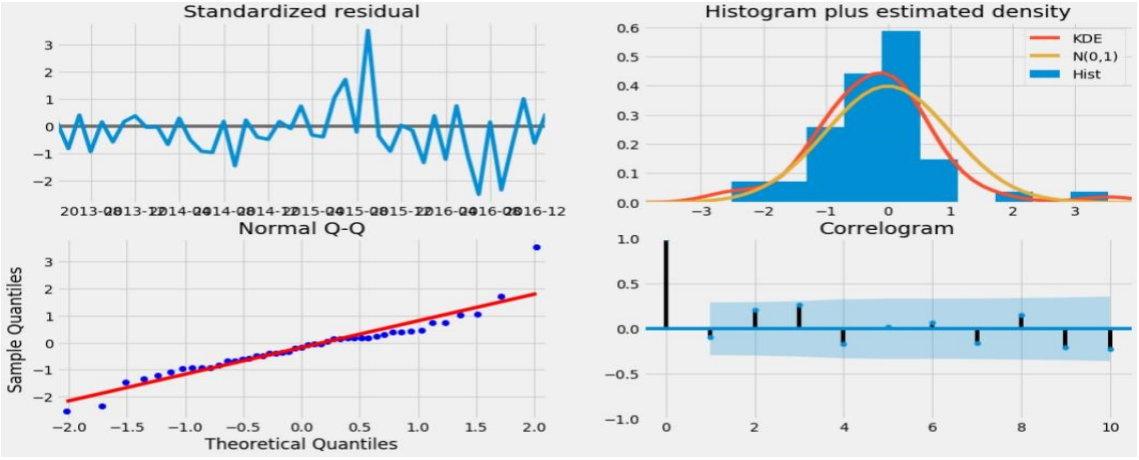

FIGURE 5.5: plots of residuals;(a) Standardized residuals over time;(b) Estimated density frequency distribution histogram; (c) Normal Q-Q plot; (d) Auto correlation

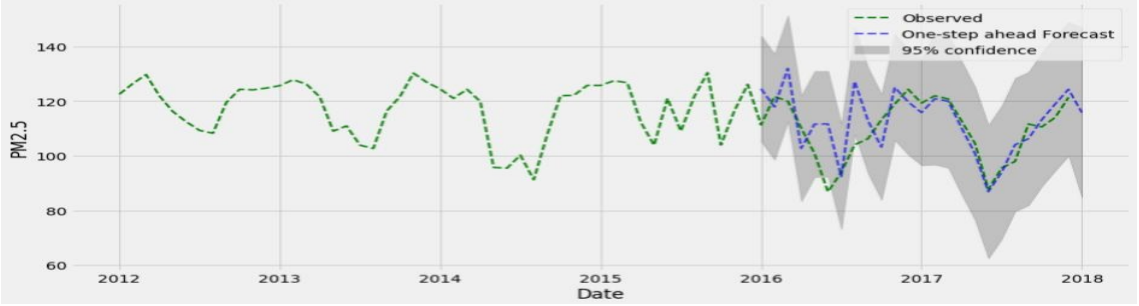

FIGURE 5.6: Comparison between the real data forecast values for one step ahead for the forecast

Figure 5.6 shows the forecasting values and one-step ahead forecasting surrounded by the confidence limits. In Figure 5.7, particulate matter ($PM_{2.5}$) observed values are from the year 2012 to 2017 as represented by the dotted sky blue which is close to the predicted next 8 years form 2018 to 2025 as shown in the dotted red line. Finally, the result of SARIMA model are acceptable for prediction of particulate matter ($PM_{2.5}$) on Chiayi city in Taiwan.

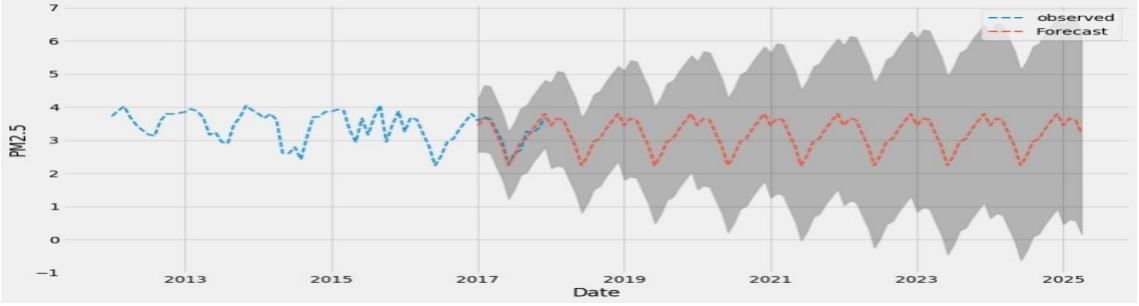

FIGURE 5.7: Forecasted values for the monthly average particulate matter ($PM_{2.5}$) from 2017 to 2025.

5.4 Conclusion

In this chapter, we have proposed SARIMA model time-series analysis used to analyze the air pollution on TAQMN data in Taiwan is presented. Particularly, there are 76 air pollution stations readings are measured by TAQMN data from 2012 to 2017. The best parameters (0, 0, 2) × (0, 1, 0, 12) for particulate matter ($PM_{2.5}$) prediction SARIMA model based on the statistical calculations such as MAPE, MAE, MPE, MSE, and RMSE. The results show that the actual values and predicted values are very close to each other. Furthermore, we predicted the next 8 years of the particulate matter ($PM_{2.5}$) levels. Finally, we conclude that SARIMA model is better for forecasting the TAQMN data. Furthermore in the future we implement the prediction particulate matter ($PM_{2.5}$) using different normalized feature engineering machine learning models.

Chapter 6

Forecasting Air Pollution Particulate Matter ($PM_{2.5}$) Using Machine Learning Models

6.1 Introduction

Air is one of the major components for all living organisms on the earth. From the last 50 years, still pollution is increasing because of urbanization, industrialization, automobiles, power plants, chemical activities, and some of the other natural activities such as agricultural burning, volcanic eruptions, and wildfires. All these activities cause the pollution growth, particularly particulate matter (PM) is one of the significant reason for air pollution [54]. Recently, air pollution turned into a serious issue on the earth and furthermore it is one of the significant reasons for death. The World Health Organization (WHO) estimated that around 6.9 millions of deaths owing to air contamination throughout the world. The recent updates of the Global Burden of Disease (GBD) study ranked the particulate matter (PM) as the 4 leading cause of death out of 85 risk factors, being responsible for over 5 million deaths in 2017.

The proposed machine learning models are used for $PM_{2.5}$ concentration prediction. We collected the data, Taiwan Air Quality Monitoring Network (TAQMN) from Taiwan country to train the

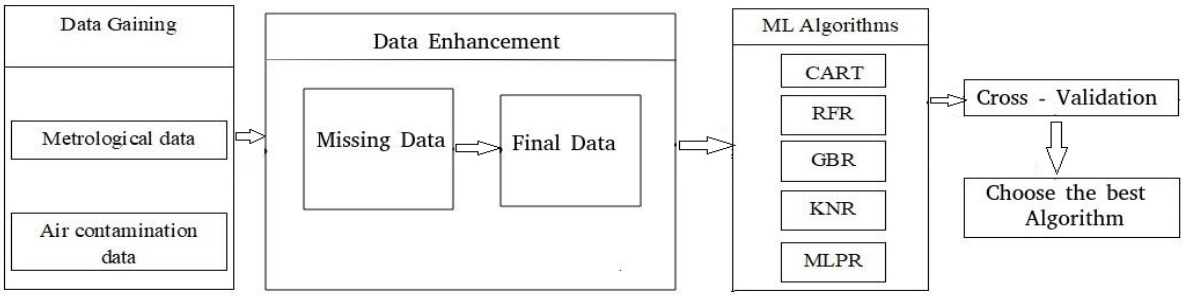

FIGURE 6.1: The proposed architecture of prediction model for air pollution.

model. TAQMN is containing air pollution data as well as meteorological data. The experiments are conducted on related data from Pingtung station, Newport city, Taiwan country.

6.2 The Proposed Model

The parameters affecting the air pollution incorporate topographic segments (x, y, z) because of the tremendous zone of Taiwan and huge tallness changes meteorological parts because of their immediate or round about impacts on poison outflow week-time segments, as in different long stretches of the week, the traffic in Taiwan is unique and in like manner the measure of toxin focus will be extraordinary [55]. For instance, on the ends of the week, Taiwan encounters less traffic thickness and the long stretch of year part additionally changes because of residue, and temperature reversal prompts changes in the degree of the toxins. Then in this examination, the air pollution estimations of the two closest air pollution station neighbors are utilized as a compelling parameter in the expectation of the air pollution.

To forecast the air pollution, we utilized machine learning strategies, such as linear regression, random forest regressor, gradient boosting regressor, k neighbors regressor, MLP regressor, and Decision Tree regressor CART [56]. Figure 6.1 shows the proposed architecture of the prediction models for air pollution.

After the preprocessing and enhancement of data, the filtered meteorological data and air pollution data are used to forecast the $PM_{2.5}$ level. In order to perform prediction, a set of machine learning algorithms used such as linear regression, random forest, gradient boosting, K-neighbors, MLP,

decision tree CART. Finally, the cross-validation are used to select the best model, that can be used for forecasting [57].

6.2.1 Random Forest

Random forest algorithm is the most popular ensemble learning model that can used for classification and regression problem. Moreover, random forest creates the forest with a number of trees, and the algorithm is more robust when the number of trees is large[58].

6.2.2 Gradient Boosting Regression

Recently, one of the most widely used ensemble algorithms is gradient boosting (XG Boost) and it is can be used for regression and classification problems. Moreover, XG Boost is designed for more speed and performance based on gradient boosted decision trees by equation 6.1.

$$D_{t+1}(i) = \frac{D_t(i)\exp\left(-\alpha_t y_i h_t(x_i)\right)}{Z_t} \tag{6.1}$$

where, $h(i)$ will be predicted by the learner, y is the actual response (+ 1 or -1), and α is the kind of learning rate [59].

6.2.3 Decision Tree Regression

Decision tree (DT) algorithm is a well known machine learning model that falls under the category of supervised learning. Moreover, it can be used for both classification and predictions tasks. The basic idea of DT is use the tree representation to solve the proposed problem in which each leaf node corresponds to a specific class label and columns are represented on the internal nodes of the created tree [60]. The main problem to apply DT is to select the attribute for the root node in each level.

6.2.4 MLP Regression

MLP is stand for a multilayer perception, which is a famous class of Artificial Neural Network (ANN). Moreover, MLP is consists of multiple layers of perceptrons or at least three layers of nodes namely input layer, hidden layer, and output layer. The nodes in each layer is called neurons and each neuron has an activation function such as Sigmoid, Softmax ,...., and ReLU [61].

6.3 Performance criteria

Some of the statistical evaluations are used to evaluate the model performance such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Square Error (MSE), and coefficient of determination (R^2). The criteria formulas are shown in below 6.2 to 6.5:

$$MAE = \frac{\sum_{i=1}^{m} |x_i - \hat{x}_i|}{m} \tag{6.2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m} (x_i - \hat{x}_i)^2}{m}} \tag{6.3}$$

$$MSE = \frac{1}{m} \sum_{j=1}^{m} (x_i - \hat{x}_i)^2 \tag{6.4}$$

where, m is the number of observations, $\hat{x}_i$ is the predicted value, and x_i is the actual value.

$$R^2 = \left[\frac{1}{M} \frac{\sum_{j=1}^{M} \left[(Y_j - \bar{Y}) \left(X_j - \bar{X} \right) \right]}{\sigma_y \sigma_x} \right]^2 \tag{6.5}$$

Where, M is the number of observations, σ_x is the standard deviation of the observation X, σ_y is the standard deviation of Y, X_j is the observed values, $\bar{X}$ is the mean of the observed values, Y_j is the calculated values, and $\bar{Y}$ is the mean of the calculated values [62].

6.4 Experimental Results

The major pollutant city is Newport that is selected and focuses only on particulate matter $PM_{2.5}$ because its the most harmful effect of pollutant in the air pollution. We collected the data in the form of hourly then it's converted into daily as well as monthly. Likewise, the proposed model gradient

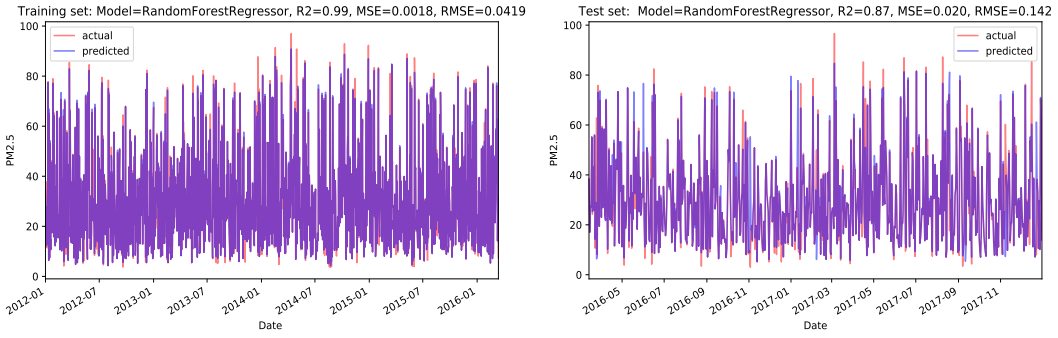

FIGURE 6.2: prediction model for Random Forest regression.

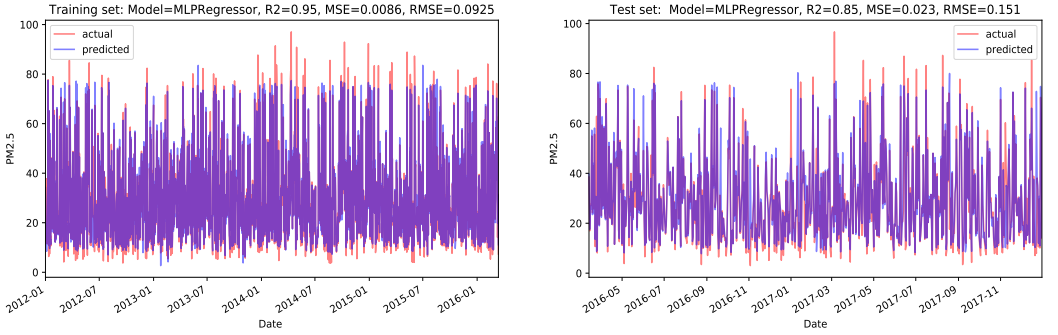

FIGURE 6.3: prediction model for MLP Regression.

TABLE 6.1: The best result of the diagnostics train and test values for different machine learning models for forecasting of the $PM_{2.5}$

	Train				**Test**			
Models	R^2	**MSE**	**RMSE**	**MAE**	R^2	**MSE**	**RMSE**	**MAE**
Linear Regression	0.7927	0.0331	0.1819	0.1186	0.7207	0.0446	0.2111	0.1364
Lasso	0.6262	0.0597	0.2443	0.1888	0.5652	0.0665	0.2579	0.1918
Ridge	0.7927	0.0331	0.1819	0.1186	0.1104	0.1361	0.3690	0.1646
Random Forest Regressor	0.9872	0.0020	0.0451	0.0125	0.8699	0.0199	0.1411	0.0372
Gradient Boosting Regressor	**0.9983**	**0.0002**	**0.0161**	**0.0101**	**0.8891**	**0.0169**	**0.1302**	**0.0380**
K Neighbors Regressor	1	0	0	0	0.6755	0.0496	0.2228	0.1491
MLP Regressor	0.9471	0.0084	0.0918	0.0388	0.8587	0.0216	0.1470	0.0562
Decision Tree Regressor CART	0.9996	6.2811	0.0079	0.0023	0.7992	0.0307	0.1753	0.0421

boosting regression was examined with including models like linear regression, lasso regression, ridge regression, random forest regression, K-neighbors regression, MLP regression, and decision tree regression. The experiments are done by using different models, and for each model the cross-validation and performance criteria is used to measure the accuracy of the model. For instance, the determination coefficient explains the relation between the actual values and predicted values.

Figure 6.2 shows the results of prediction using random forest regression algorithm, left part

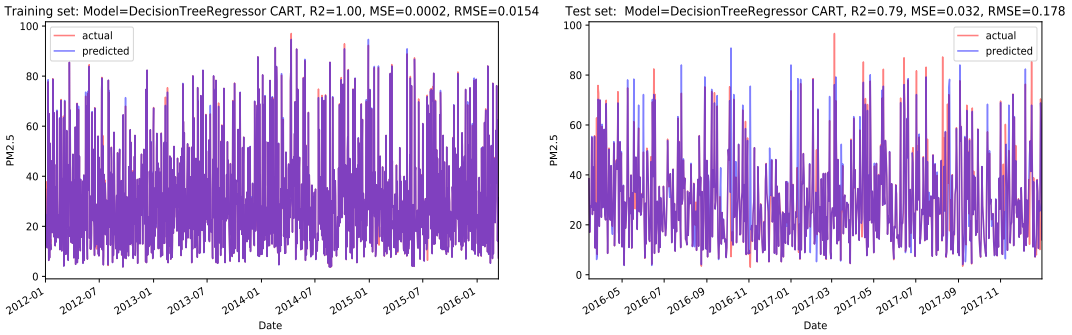

FIGURE 6.4: prediction model for Decision Tree Regression.

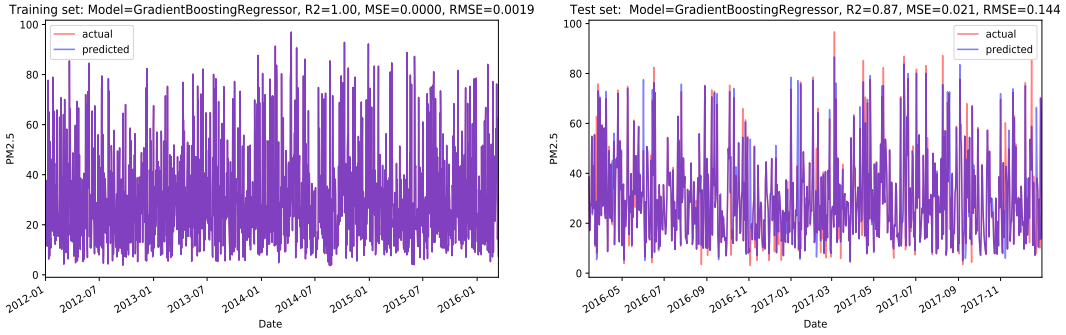

FIGURE 6.5: prediction model for Gradient Boosting Regression.

TABLE 6.2: The comparison results for proposed with model others models for forecasting of the $PM_{2.5}$

Models	R^2	**RMSE**
Bayesian Down-scaling model MCMC algorithms	0.7062	0.6895
LSTM(L)	0.7485	0.4236
CART-CM-LF-AN	0.7909	0.3258
Gradient Booting regression	**0.9336**	**0.1302**

indicate the training data and right part for testing data where R^2 is 0.99 and 0.87 respectively. Moreover, the red color line indicates the actual values and blue color line indicates the prediction of the particulate matter $PM_{2.5}$. Similarly, Figure 6.3 shows the results of prediction for MLP regression algorithms, R^2 is 0.95 and 0.85 for training data and testing data respectively. Figure 6.4 shows decision tree algorithm training data and testing data R^2 are 1.00 and 0.79 respectively. Figure 6.5 shows the results of prediction using gradient boosting regression algorithm, R^2 the training data and testing data are 1.00 and 0.89 respectively. Moreover, coefficient of determination values are shown in Table 6.1.

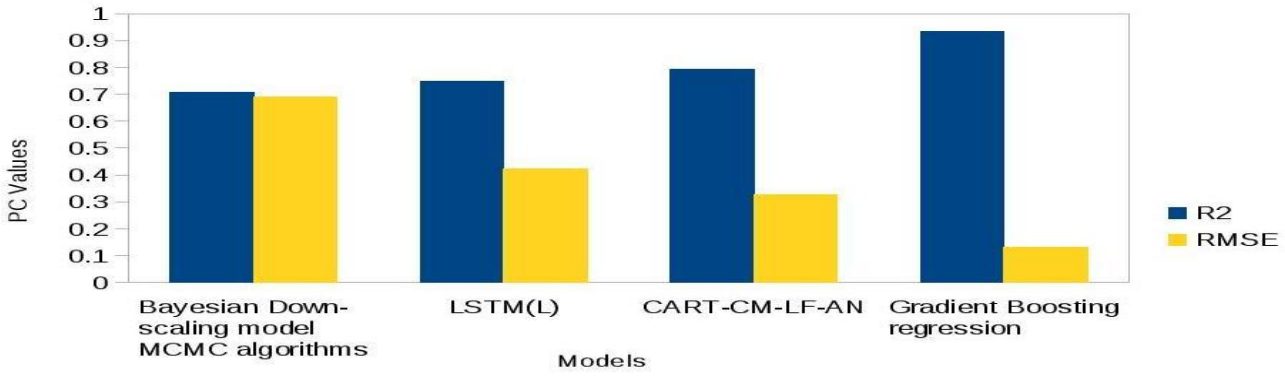

FIGURE 6.6: The comparison results for proposed with model others models for forecasting of the $PM_{2.5}$

Table 6.1 show the results for each algorithm, gradient boosting regression performs better compared with the rest of algorithms. Moreover, Table 6.2 and Figure 6.6 presents the comparison between the best model from Table 6.1 and the related work such as Bayesian-Down-scaling model, LSTM, and CART-CM-LF-AN.

6.5 Conclusion

In this work, the proposed machine learning models to analyze the air pollution on TAQMN data in Taiwan is presented. Particularly, there are 76 air pollution stations are recorded in TAQMN data from 2012 to 2017. The particulate matter $PM_{2.5}$ prediction are done using machine learning models based on the statistical calculations of metrics such as MAE, MSE, RMSE, and R^2. The results show that the proposed models values are perform better compares to the previous models and also it show that the actual values and predicted values are very close to each other. Finally, we conclude that gradient boosting regressor model is better for predicting the air pollution on the TAQMN data.

Chapter 7

Spatio-Temporal Clustering Analysis for Air Pollution Particulate Matter ($PM_{2.5}$) Using a Deep Learning Model

7.1 Introduction

Air contamination is an overall natural issue that can prompt critical biological and ecological impacts and undermine human well being, particularly the barometric fine particulate matter ($PM_{2.5}$) in numerous Asian urban areas [63]. It is important that the greater part of these passing occurred in creating nations, especially in China, Taiwan, and India. Particulates, particularly $PM_{2.5}$ (vaporized particles with the streamlined distance across under less than 2.5 microns in diameter), are the deadliest type of air contamination's. Epidemiological investigations have indicated that presentations to $PM_{2.5}$ are related to expanded cardiovascular and respiratory bleakness and mortality. With the quick improvement of urbanization and industrialization, $PM_{2.5}$ contamination's in Taiwan, particularly in the North part, is getting genuine, which can be seen by sensors. Subsequently, the examination of the spatio-temporal example of $PM_{2.5}$ across Taiwan is basic for doing ecological epidemiological investigations [64].

The primary targets were:

1. To investigate the spatial and Taiwan dispersion qualities of $PM_{2.5}$ in various geological regions and under the differed natural administration.

2. To impact components of air contamination source, monetary, geographical, and meteorological elements on $PM_{2.5}$ contamination designs.

7.2 Methods

The fundamental techniques used in this paper to examine the dissemination of $PM_{2.5}$ and its persuasive elements in the various urban areas are group investigation and unaided calculation examination. Cluster investigation, which is likewise called cluster examination, is a sort of technique that reviews the order of tests or records depending on their own highlights. The alleged "bunch", is an assortment of comparable components [65]. The standard of clustering examination is to analyze the highlights of components straightforwardly and put the components with comparable characteristics into one grouping, and components with various characteristics into various arrangements.

The principle steps of clustering investigation based on python can be coordinated as follows:

Stage 1: Find the likenesses and dissimilarities between various factors in the informational collection (the checking information of various locations in this paper).

Stage 2: Define the linkage between various factors utilizing the "linkage" work.

Stage 3: Assess the previous count impact dependent on the "cophenetic" work.

Stage 4: Establish various groups through the "cluster" work.

Stage 5: Obtain the heatmap through "plot".

7.2.1 Autoencoder

Autoencoder is a kind of neural organization with a balanced structure from encoding to deciphering layers and a similar number of information and yield factors in an unaided or regulated way [66]. It intends to become familiar with a proficient information portrayal (encoding) regularly for dimensional decrease. Figure 7.1 presents a run of the mill autoencoder with input, yield, coding, and shrouded layers.

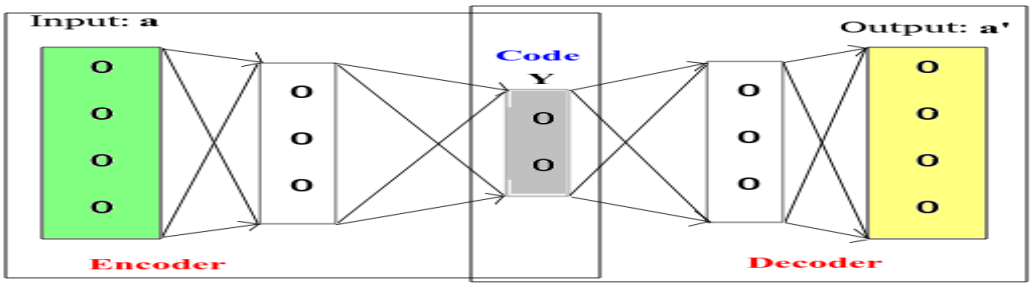

FIGURE 7.1: An ordinary autoencoder with a symmetrical organize topology and the same input and output.

7.2.2 Residual learning in CNN

Residual profound CNN has been proposed to handle the issue of precision corruption in more profound learning organizations[67]. In this design, every leftover unit incorporates a few (for example two) constant convolutional layers with cluster normalization's and ReLU actuations; all the lingering units are stacked consistently to expand the profundity and the speculation capacity of the organization.

7.2.3 Spatial autoencoder

Spatial autoencoder exists in most geoscience wonders for surface estimations. Profound learning systems can not straightforwardly implant the structure of spatial autoencoder. But spatial autoencoder can be captured in profound arrange relapse by utilizing spatial arranges and their subsidiaries (like square and item) as ones of the co-verities. We may moreover utilize the approach of closest neighbors or part thickness to infer the layer of spatial closest neighbors [68]. Spatial autoencoder, in case inserted inside the models, can make strides in the model's execution.

7.2.4 Autoencoder-based architecture

Autoencoder (Figure 7.2) is a reflected arrangement with encoding and interpreting layers. For the interpreting layers, each covered up layer may have a diverse number of hubs that present the variety of compression or adjustment of the measurements.

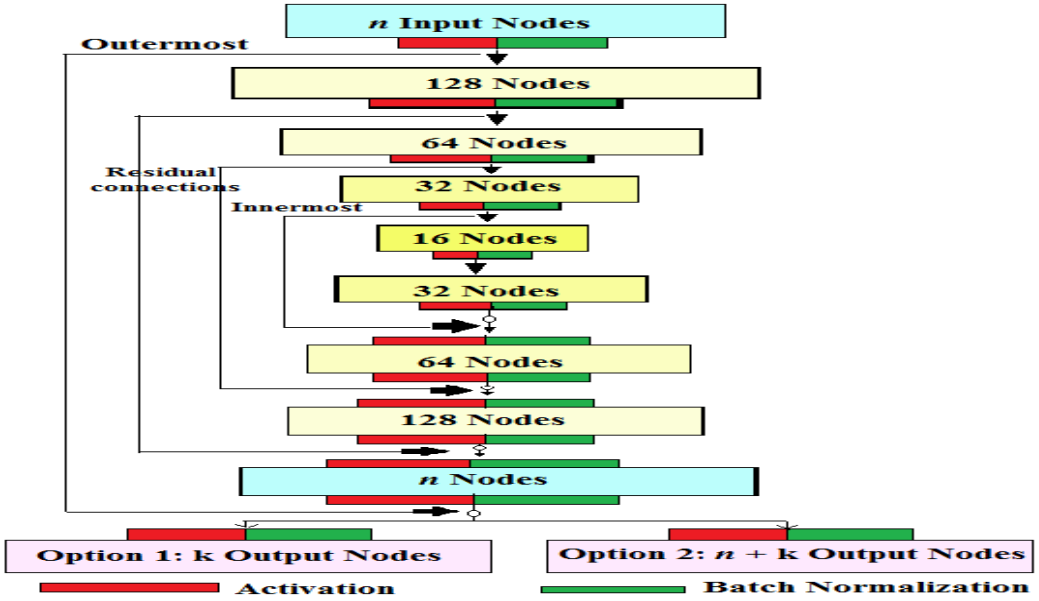

FIGURE 7.2: The architecture of an ordinary autoencoder based remaining deep arrange (n input, n output with symmetrical covered up layers with changes of the nodes number)

7.2.5 Residual identity association and implementation

We utilize the easy route association of personality mapping from the encoding layer to its comparing reflected interpreting layer to actualize remaining systems in a nested way [69] from the furthest to deepest layers (Figure 7.2). Depending on whether not one or the other enhancement nor group normalization, fair actuation, or both actuation and bunch normalization are included after the yields of the covered up layers

Accept the center layer M, the decoding layer l, and it reflects the interpreting layer L. Given xl and gl to be input and output of the encoding layer l, separately, and xL and gL to be input and output of the decoding layer L, individually. With the expansion of the remaining character association, we have the equations 7.1 to 7.3 :

$$g_L = a_l + F\left(a_L, W_{1L}\right) \tag{7.1}$$

$$a_{L+1} = f\left(g_L\right) \tag{7.2}$$

$$g_L = a_l + F\left(\cdots f\left(a_L, W_{1l}\right)\cdots\right) \tag{7.3}$$

$$\begin{aligned}\frac{\partial L}{\partial a_l} &= \frac{\partial L}{\partial f_L\left(g_L\right)} \frac{\partial f_L\left(g_L\right)}{\partial g_L} \frac{\partial g_L}{\partial a_l} \\ &= \frac{\partial L}{\partial f_L\left(g_L\right)} \frac{\partial f_L\left(g_L\right)}{\partial g_L} \frac{\partial g_L}{\partial a_l}\left(1+\frac{\partial}{\partial a_l} F\left(\cdots f_l\left(a_l, W_{1l}\right)\cdot\right)\right)\end{aligned} \tag{7.4}$$

Where $f_L\left(g_L\right)$ speaks to the actuation function for g_L. In case we utilize the remaining interface of option 1 in Figure **??**-a (no actuation after an expansion of the alternate route character association), we are able to get the easier adaptation for equation7.4:

$$\frac{\partial L}{\partial a_l} = \frac{\partial L}{\partial g_L} \frac{\partial g_L}{\partial a_l} = \frac{\partial L}{\partial g_L}\left(1+\frac{\partial}{\partial a_l} F\left(\cdots f_l\left(a_l, W_{1l}\right)\cdot\right)\right) \tag{7.5}$$

For choice 2 and 3 in Fig. **??**(a), if the enactment capacity of ReLU or exponential linear unit (ELU) or linear unit (LU), or/and clump standardization are added [$f_L\left(g_L\right)$ in (equation 7)] for better nonlinear demonstrating, they can likewise ensure the pleasant property previously mentioned by (equation 7.1).

The autoencoder structure gives fair network geography so the shallow layers have their reflected partners of more profound layers in the unraveling segment. Accordingly, the autoencoder-based structure is a characteristic alternative for lingering relapse profound networks in Figure 7.3.

7.3 Results and Discussion

Furthermore, Figure 7.3 show the clustering convolution autoencoder fully connected the layer's loads speak to the group centred which can be instated via preparing a K-means.

There are four models can be represented in the in cluster analysis with the different stations first load the clustering model trained weights, second as Build the clustering model with a convolutional autoencoder, third as Model to train clustering and autoencoder same time (Convolutional), and last one as Model to train clustering and autoencoder same time (Fully connected) for evaluation of the clustering performance as Accuracy (Acc), Normalized mutual information (Nmi), Adjusted rand index (Ari), and loss. For all the models have to be applying to getting the performance evaluation as shown in Table 7.1.

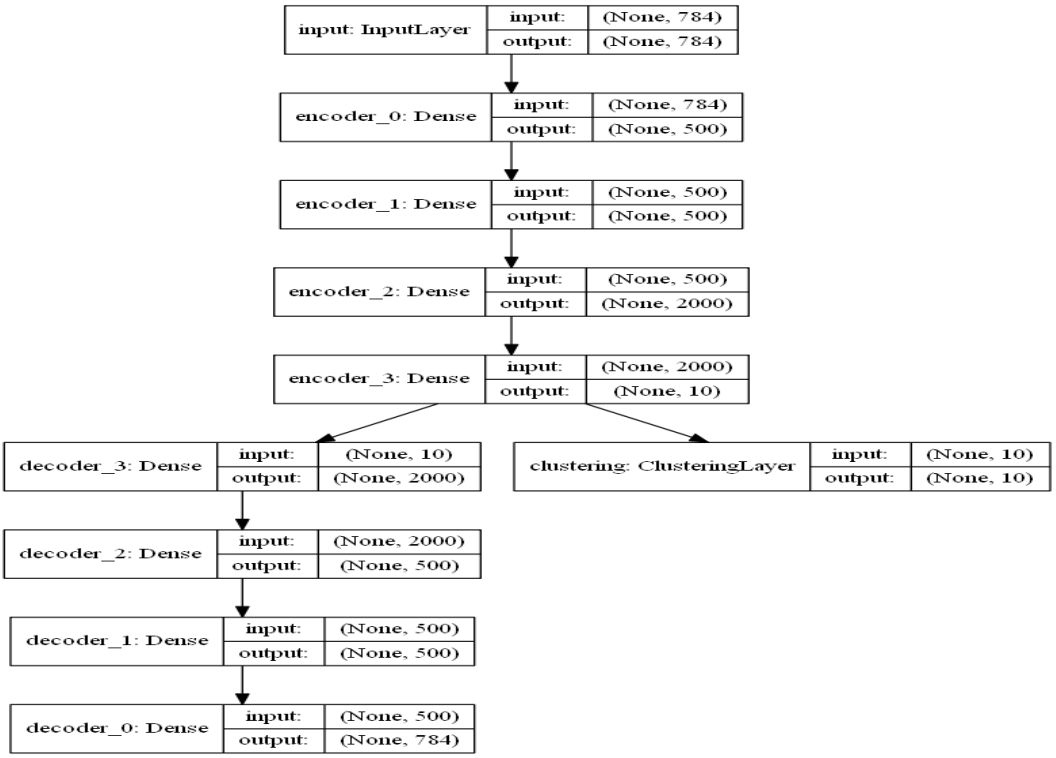

FIGURE 7.3: Clustering and Convolutional autoencoder fully connected.

TABLE 7.1: Evaluation of the clustering performance

Models	Acc	Nmi	Ari	Loss
Load the clustering model trained weights	0.96	0.91	0.92	0.75
Build the clustering model with a convolutional autoencoder	0.80	0.77	0.69	0.37
Model to train clustering and autoencoder at the same time (Convolutional)	0.83	0.74	0.81	0.31
Model to train clustering and autoencoder at the same time (Fully connected)	0.97	0.90	0.91	0.23

As expected to be productively figured by the Hungarian calculation and For all models heat-map of the clustering as shown in Figure 7.4 (a,b,c,d) can be represented a) Partition clustering model confusion matrix. b) Convolutional autoencoder confusion matrix. c) Clustering and convolutional autoencoder confusion matrix. d) Clustering and convolutional autoencoder fully connected confusion matrix.

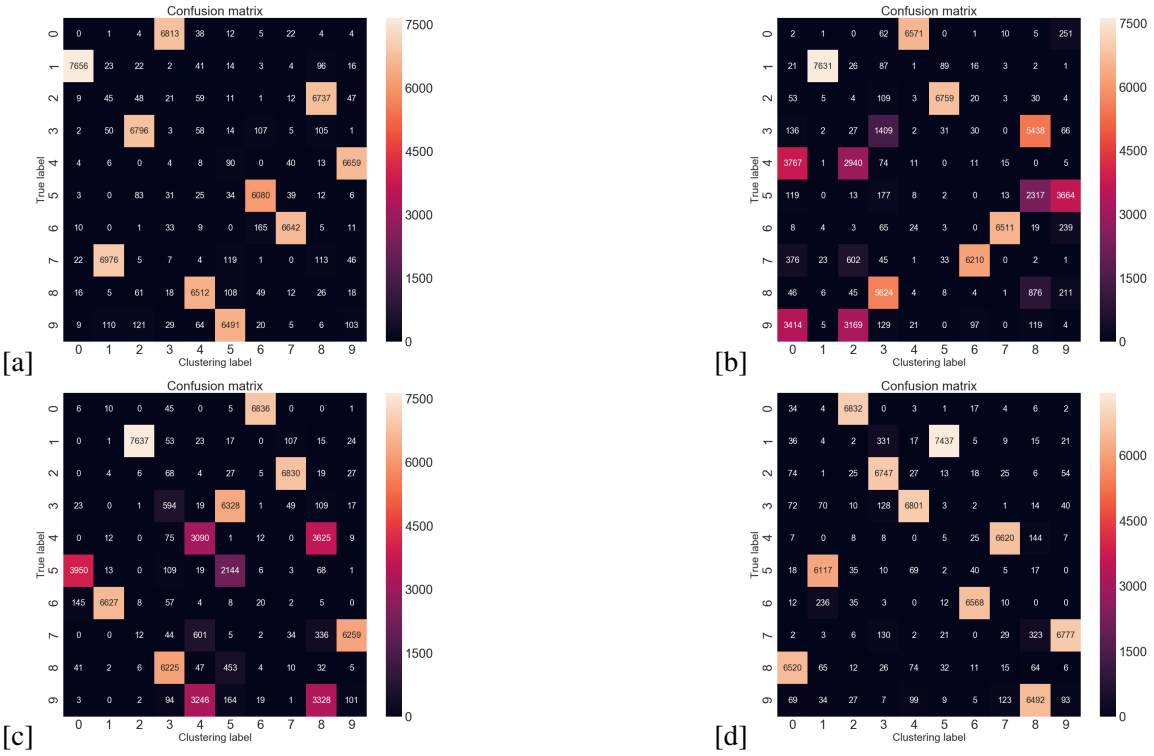

FIGURE 7.4: a) Partition clustering model confusion matrix. b) Convolutional autoencoder confusion matrix. c) Clustering and Convolutional autoencoder confusion matrix. d) Clustering and Convolutional autoencoder fully connected confusion matrix.

7.4 Conclusion

In the conclusion the proposed deep learning models to analyze the particulate matter $PM_{2.5}$ air pollution on TAQMN data in Taiwan nation to construct a model to perform clustering examination with unlabeled informational collections. Pre-prepared autoencoder assumed a critical function in the dimensional decrease and boundary in the statement, at that point exclusively constructed clustering layer was prepared against an objective appropriation to refine the precision further. There are four models can be represented in the cluster analysis with the different stations first load the clustering model trained weights, second as building the clustering model with a convolutional autoencoder, third as a model to prepare grouping and autoencoder simultaneously (convolutional), and the last one as a model to prepare clustering and autoencoder same time (completely associated) for evaluation of the clustering performance as accuracy (Acc), normalized mutual information (Nmi), adjusted rand index (Ari), and loss. In all the above models fully connected model is a better result performance model.

Chapter 8

Conclusions and Future Directions

In this chapter, we will present the whole thesis objectives and results. We will display the scientific contributions that this research work made. In the end, we will introduce some points for the future work of this research. The work conclusions are introduced in section 8.1 and the future research direction is presented in section 8.2.

8.1 Conclusions

Air contamination assumes a significant part in our regular routine, and without meteorologists and forecasters, we would experience issues arranging our everyday exercises. As should be obvious, air contamination is anything but a basic subject like we might have been thinking. The investigation of the diverse peculiarity requires the utilization of science, math, and various kinds of gear and innovation, and information. Indeed, even with every one of these hardware, information, and perception apparatuses, air contamination keeps on being a subject to contemplate in light of the fact that it is continually evolving. Meteorologists and forecasters anticipate air contamination and its potential changes, yet truly, air contamination is as yet unusual.

A writing audit of this thesis is done to investigate and recognize the significance, and difficulties of air contamination information examination. Just as, checking on various procedures like forecast, and clustering for the enormous informational collections which assisted us with achieving this work.

The First part of this work we developing the Taiwan country air pollution data warehouse, the multidimensional data model is used to implementation of the data warehouse its collection of data cube, data mart, and meta data using star schema and snowflake schema the statistical analysis of five years of data of Taiwan country all stations are statistical evaluation such as minimum, maximum, mean, standard deviation, and coefficient variation the descriptive statistics values determining the level of pollutant emissions. In selected stations are considered for resolving decision making issues, it can be easy to access the data and size of the memory can be decreased.

Through the experimental results of this thesis, it demonstrates the possibility of we outlined the concept of investigating the phenomenon of time series of correlation analysis of the particulate matter ($PM_{2.5}$) with selected cities by using the visualization framework with the cross-correlation analysis method we analyzed the Annan, Newport, Good, and McMug cities are positive and a negative correlation of the air pollution pollutants and meteorological factors. Some cities are showing the maximum positive correlation and less a negative correlation. Visualization framework and a cross-correlation method determine the particulate matter ($PM_{2.5}$) and wind speed for selected cities Annan, Newport, Good, and McMug first two cities are a positive correlation and the other two cities are a negative correlation, then Pearson's Correlation r and p for particulate matter ($PM_{2.5}$) with wind speed to the different stations.

The anticipating of traditional techniques might be inadequate to gauge future qualities effectively as a result of the deficiency of example acknowledgment of ARIMA and the issue of slack assurance. We have proposed SARIMA model time-series analysis used to analyze the air pollution on TAQMN data in Taiwan is presented. Particularly, there are 76 air pollution stations readings are measured by TAQMN data from 2012 to 2017. The best parameters (0, 0, 2) × (0, 1, 0, 12) for particulate matter ($PM_{2.5}$) prediction SARIMA model based on the statistical calculations such as MAPE, MAE, MPE, MSE, and RMSE. The results show that the actual values and predicted values are very close to each other. Furthermore, we predicted the next 8 years of the particulate matter ($PM_{2.5}$) levels. Finally, we conclude that SARIMA model is better for forecasting the TAQMN data.

In this work, the proposed machine learning models to analyze air pollution on TAQMN data in Taiwan are presented. The particulate matter $PM_{2.5}$ predictions are done using machine learning models based on the statistical calculations of metrics such as MAE, MSE, RMSE, and R^2. The

results show that the proposed model's values perform better compares to the other machine learning models such as linear regression, lasso regression, ridge regression, random forest regression, K-neighbors regression, MLP regression, and decision tree regression. The experiments are done by using different models, and for each model, the cross-validation and performance criteria are used to measure the accuracy of the model. And also previous research models such as Bayesian down-scaling model MCMC algorithm, LSTM(L), and CART-CM-LF-AN for an instance, the determination coefficient explains the relation between the actual values and predicted values. and also it shows that the actual values and predicted values are very close to each other. Finally, we conclude that the gradient boosting regressor model is better for predicting air pollution on the TAQMN data.

Finally, the proposed deep learning models to analyze the particulate matter $PM_{2.5}$ air pollution on TAQMN data in Taiwan nation to construct a model to perform clustering examination with unlabeled informational collections. The pre-prepared autoencoder assumed a critical function in the dimensional decrease and boundary in the statement, at that point exclusively constructed clustering layer was prepared against an objective appropriation to refine the precision further. There are four models that can be represented in the cluster analysis with the different stations first load the clustering model trained weights, second as building the clustering model with a convolutional autoencoder, third as a model to prepare grouping and autoencoder simultaneously (convolutional), and the last one as a model to prepare clustering and autoencoder same time (completely associated) for evaluation of the clustering performance as accuracy (Acc), normalized mutual information (Nmi), adjusted rand index (Ari), and loss. In all the above models fully connected model is a better result performance model.

8.2 Open Problems and Future Works

Albeit the proposed models and examination of air contamination information, there are a few upgrades that can, in any case, be made. Since these models showed better work for the prediction and achieve amazing outcomes. The work done in this examination leaves various open inquiries and some exploration focuses which can be tended to later on, such as:

1. The two fundamental issues of artificial neural organizations and profound learning calculations are the gigantic quantities of boundaries and the investigation of their designs and structures. Much still needs to be done to choose the best setups and designs for deep learning organizations.

2. Advancement: Much still needs to be improved comprehend the victories and disappointments of preparing profound designs, both in the regulated and unaided cases. Despite the fact that regularization impacts can be significant on little datasets, the impacts that continue on exceptionally enormous datasets propose some improvement issues are involved.

3. Plan of profound models to gain from less preparing information: Sometimes just a predetermined number of preparing information is accessible, so incredible models are needed to accomplish an upgraded taking in capacity from restricted information.

4. Reasonable preprocessing procedures can likewise be concocted to build expectation precision. Additionally, the expectation execution can be considered by the incorporation of covariates, utilizing the multivariate models, and improving the model boundaries.

5. Execution of deep learning calculations on cell phones: It ought to be noticed that deep learning approaches as a rule require great computational time. Subsequently, we can run AI calculations in cell phones with lower power utilization. Besides, air pollution determination can be anticipated proficiently utilizing portable GPU devices.

6. The proposed models can be more dependable and more proficient by utilizing similar TAQMN informational collection to run the tests and crossing over 20 years. Additionally for greater improvement of this work can be done utilizing enhancement procedures, for example, framework scan that consequently picks the best qualities for boundaries of the models.

List of Publication

International Journals

- Doreswamy, Harishkumar, K. S: **Paper Title**,– **"Visualization of Correlation for Air Pollution Time Series using TAQMN Data."** *International Journal of Recent Technology and Engineering (IJRTE)*, (ISSN:2277-3878(Online), Volume=8 Issue-3, September-2019 page No.:7922-7927).

- Doreswamy, Harishkumar, K. S., K. M. Yogesh, and Ibrahim Gad: **Paper Title**,–**"Forecasting air pollution particulate matter ($PM_{2.5}$) using machine learning regression models."** *Procedia Computer Science*, 171 (2020): pp 2057-2066.

- Doreswamy, Harishkumar, K. S., and Ibrahim Gad: **Paper Title**,–**"Time Series Analysis for Prediction of $PM_{2.5}$ Using Seasonal Autoregressive Integrated Moving Average (SARIMA) Model on Taiwan Air Quality Monitoring Network Data."** *Journal of Computational and Theoretical Nanoscience*, 17.9-10 (2020): 3964-3969.

International Conferences

- Doreswamy, Harishkumar, K. S:, **Paper Title**,– **"A survey of air pollution studies using data mining techniques in smart city."**,– *National Conference on Science, Engineering And Management (NCSEM 2018)*, IEEE Proceedings Print (ISSN 2321-2241).

- Doreswamy, Harishkumar, K. S:, **Paper Title**,– **" Multidimensional Data Model for Air Pollution Data Analysis"**.,– *IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI 2018)*, pp. 1684-1689.

- Doreswamy, Harishkumar, K. S, Ibrahim Gad, and Yogesh K. M:, **Paper Title**,– **" Spatio temporal Clustering Analysis for Air Pollution Particulate Matter ($PM_{2.5}$) Using a Deep Learning Model"**.,– *IEEE International Conference on Computing, Communication, and Intelligent Systems (ICCCIS 2021)*, (pp. 529-535).

Bibliography

[1] Alexandrov, Alexander, Rico Bergmann, Stephan Ewen, Johann-Christoph Freytag, Fabian Hueske, Arvid Heise, Odej Kao et al. "The stratosphere platform for big data analytics." The VLDB Journal 23, no. 6 (2014): 939-964.

[2] Akaraci, Selin, Muhammad Arslan Usman, Muhammad Rehan Usman, and Dong Joon Ahn. "From smart to smarter cities: Bridging the dimensions of technology and urban planning." In 2016 International Conference on Smart Green Technology in Electrical and Information Systems (ICSGTEIS), pp. 74-78. IEEE, 2016.

[3] Kumar, A., and Anand Prakash. "The role of big data and analytics in smart cities." Int J Sci Res (IJSR) 6, no. 14 (2014): 12-23.

[4] Smith, William H. Air pollution and forests: interactions between air contaminants and forest ecosystems. Springer Science and Business Media, 2012.

[5] Hashem, IbrahimAbakerTargio, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. "The rise of “big data” on cloud computing: Review and open research issues." Information systems 47 (2015): 98-115.

[6] França, Katlein, Aparecida Porto França, and Reginaldo de França. "Environmental psychodermatology: stress, environment and skin." In Stress and skin disorders, pp. 47-53. Springer, Cham, 2017.

[7] Kerner, Boris S., Peter Hemmerle, Micha Koller, Gerhard Hermanns, Sergey L. Klenov, Hubert Rehborn, and Michael Schreckenberg. "Empirical synchronized flow in oversaturated city traffic." Physical Review E 90, no. 3 (2014): 032810.

[8] Manogaran, Gunasekaran, Chandu Thota, Daphne Lopez, V. Vijayakumar, Kaja M. Abbas, and Revathi Sundarsekar. "Big data knowledge system in healthcare." In Internet of things and big data technologies for next generation healthcare, pp. 133-157. Springer, Cham, 2017.

[9] Godish, Thad, Wayne T. Davis, and Joshua S. Fu. Air quality. CRC Press, 2014.

[10] Kitchin, Rob. "The real-time city? Big data and smart urbanism." GeoJournal 79, no. 1 (2014): 1-14.

[11] Bellinger, Colin, Mohomed Shazan Mohomed Jabbar, Osmar Zaïane, and Alvaro Osornio-Vargas. "A systematic review of data mining and machine learning for air pollution epidemiology." BMC public health 17, no. 1 (2017): 1-19.

[12] Ding, Weifu, Jiangshe Zhang, and Yee Leung. "Prediction of air pollutant concentration based on sparse response back-propagation training feedforward neural networks." Environmental Science and Pollution Research 23, no. 19 (2016): 19481-19494.

[13] Zhang, Jiangshe, and Weifu Ding. "Prediction of air pollutants concentration based on an extreme learning machine: the case of Hong Kong." International journal of environmental research and public health 14, no. 2 (2017): 114.

[14] van der Zee, Saskia C., Paul H. Fischer, and Gerard Hoek. "Air pollution in perspective: Health risks of air pollution expressed in equivalent numbers of passively smoked cigarettes." Environmental research 148 (2016): 475-483.

[15] Huang, Keyong, Qingyang Xiao, Xia Meng, Guannan Geng, Yujie Wang, Alexei Lyapustin, Dongfeng Gu, and Yang Liu. "Predicting monthly high-resolution PM2. 5 concentrations with random forest model in the North China Plain." Environmental pollution 242 (2018): 675-683.

[16] Jung, Chau-Ren, Bing-Fang Hwang, and Wei-Ting Chen. "Incorporating long-term satellite-based aerosol optical depth, localized land use data, and meteorological variables to estimate ground-level PM2. 5 concentrations in Taiwan from 2005 to 2015." Environmental pollution 237 (2018): 1000-1010.

[17] Stafoggia, Massimo, Tom Bellander, Simone Bucci, Marina Davoli, Kees De Hoogh, Francesca De'Donato, Claudio Gariazzo et al. "Estimation of daily PM10 and PM2. 5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model." Environment international 124 (2019): 170-179.

[18] Awan, Faraz Malik, Roberto Minerva, and Noel Crespi. "Improving Road Traffic Forecasting Using Air Pollution and Atmospheric Data: Experiments Based on LSTM Recurrent Neural Networks." Sensors 20, no. 13 (2020): 3749.

[19] Koo, Jian Wei, Shin Wee Wong, Ganeshsree Selvachandran, and Hoang Viet Long. "Prediction of Air Pollution Index in Kuala Lumpur using fuzzy time series and statistical models." Air Quality, Atmosphere Health 13, no. 1 (2020): 77-88.

[20] Shafii, Nor Hayati Binti, Rohana Alias, Nur Fithrinnissaa Zamani, and Nur Fatihah Fauzi. "Forecasting of air pollution index PM2. 5 using support vector machine (SVM)." Journal of Computing Research and Innovation 5, no. 3 (2020): 43-53.

[21] Chang, Yue-Shan, Satheesh Abimannan, Hsin-Ta Chiao, Chi-Yeh Lin, and Yo-Ping Huang. "An ensemble learning based hybrid model and framework for air pollution forecasting." Environmental Science and Pollution Research 27, no. 30 (2020): 38155-38168.

[22] Liu, Duen-Ren, Shin-Jye Lee, Yang Huang, and Chien-Ju Chiu. "Air pollution forecasting based on attention-based LSTM neural network and ensemble learning." Expert Systems 37, no. 3 (2020): e12511.

[23] Alyousifi, Yousif, Mahmod Othman, Ibrahima Faye, Rajalingam Sokkalingam, and Petronio CL Silva. "Markov Weighted Fuzzy Time-Series Model Based on an Optimum Partition Method for Forecasting Air Pollution." International Journal of Fuzzy Systems 22 (2020): 1468-1486.

[24] Xu, Yinan, Hui Liu, and Zhu Duan. "A novel hybrid model for multi-step daily AQI forecasting driven by air pollution big data." Air Quality, Atmosphere Health 13, no. 2 (2020): 197-207.

[25] Leedy, Paul D., and Jeanne Ellis Ormrod. "Practical research: Planning and design, global edition." England: Pearson Education Limited (2015).

[26] Herman, M., S. Rivera, S. Mills, J. Sullivan, P. Guerra, A. Cosmas, and M. Kim. "The field guide to data science." Booz Allen Hamilton (2013).

[27] Kalidindi, Surya R., and Marc De Graef. "Materials data science: current status and future outlook." Annual Review of Materials Research 45 (2015): 171-193.

[28] Chen, Sheng-Po, Chih-Chung Chang, Jyh-Jian Liu, Charles C-K. Chou, Julius S. Chang, and Jia-Lin Wang. "Recent improvement in air quality as evidenced by the island-wide monitoring network in Taiwan." Atmospheric environment 96 (2014): 70-77.

[29] Mesjasz-Lech, Agata. "Urban air pollution challenge for green logistics." Transportation Research Procedia 16 (2016): 355-365.

[30] Chen, De qing, Wen-yu Wang, and Hai-kun Yang. "Application research on data warehouse of hydrological data comprehensive analysis." In 2010 3rd International Conference on Computer Science and Information Technology, vol. 9, pp. 140-143. IEEE, 2010.

[31] Gad, Ibrahim, and B. R. Manjunatha. "Hybrid data warehouse model for climate big data analysis." In 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT), pp. 1-9. IEEE, 2017.

[32] Ansari, Mohammadnaser, and Jeffrey S. Smith. "Warehouse operations data structure (wods): A data structure developed for warehouse operations modeling." Computers and Industrial Engineering 112 (2017): 11-19.

[33] Boulil, Kamal, Florence Le Ber, Sandro Bimonte, Corinne Grac, and Flavie Cernesson. "Multidimensional modeling and analysis of large and complex watercourse data: an OLAP-based solution." Ecological informatics 24 (2014): 90-106.

[34] Gui, Hao, and Mark Roantree. "A data cube model for analysis of high volumes of ambient data." Procedia Computer Science 10 (2012): 94-101.

[35] Mosavi, Amir, Sina Ardabili, and Annamaria R. Varkonyi-Koczy. "List of deep learning models." In International Conference on Global Research and Education, pp. 202-214. Springer, Cham, 2019.

[36] Liu, Hongsheng, Guofei Ren, Haoyu Chen, Qi Liu, Yingjuan Yang, and Qi Zhao. "Predicting lncRNA–miRNA interactions based on logistic matrix factorization with neighborhood regularized." Knowledge-Based Systems 191 (2020): 105261.

[37] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

[38] Han, Jiawei, and Micheline Kamber. "Data Mining: Concepts and Techniques, The Morgan Kaufinann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers." (2000): 279-28.

[39] Ariyachandra, Thilini, and Hugh Watson. "Key organizational factors in data warehouse architecture selection." Decision support systems 49.2 (2010): 200-212.

[40] Hou, Rui Lian, Ai Mei Dong, and Xiu Fang Li. "The Design and Implementation of Monitoring Data Warehouse of Ambient Air Quality." Advanced Materials Research. Vol. 424. Trans Tech Publications Ltd, 2012.

[41] Hurley, Peter J., William L. Physick, and Ashok K. Luhar. "TAPM: a practical approach to prognostic meteorological and air pollution modelling." Environmental Modelling Software 20.6 (2005): 737-752.

[42] Huang, Fangfang, Xia Li, Chao Wang, Qin Xu, Wei Wang, Yanxia Luo, Lixin Tao et al. "PM 2.5 spatiotemporal variations and the relationship with meteorological factors during 2013-2014 in Beijing, China." PloS one 10, no. 11 (2015): e0141642.

[43] Yang, Yong, and George Christakos. "Spatiotemporal characterization of ambient PM2. 5 concentrations in Shandong province (China)." Environmental science technology 49, no. 22 (2015): 13431-13438.

[44] Pearce, John L., Jason Beringer, Neville Nicholls, Rob J. Hyndman, and Nigel J. Tapper. "Quantifying the influence of local meteorology on air quality using generalized additive models." Atmospheric Environment 45, no. 6 (2011): 1328-1336.

[45] Liu, Yong, et al. "Research on the relationship between urban form and urban smog in China." Environment and Planning B: Urban Analytics and City Science 44.2 (2017): 328-342.

[46] Jorgensen, Mackenzie, et al. "An interactive big data processing/visualization framework." 2017 IEEE MIT Undergraduate Research Technology Conference (URTC). IEEE, 2017.

[47] Lu, Wei, et al. "An interactive web mapping visualization of urban air quality monitoring data of China." Atmosphere 8.8 (2017): 148.

[48] Mahmud, Shahid, Rahat Iqbal, and Faiyaz Doctor. "Cloud enabled data analytics and visualization framework for health-shocks prediction." Future Generation Computer Systems 65 (2016): 169-181.

[49] Jiang, Ping, Qingli Dong, and Peizhi Li. "A novel hybrid strategy for PM2. 5 concentration analysis and prediction." Journal of environmental management 196 (2017): 443-457.

[50] Cortina–Januchs, Maria Guadalupe, Joel Quintanilla–Dominguez, Antonio Vega–Corona, and Diego Andina. "Development of a model for forecasting of PM10 concentrations in Salamanca, Mexico." Atmospheric Pollution Research 6, no. 4 (2015): 626-634.

[51] Salazar, Ledys, Orietta Nicolis, Fabrizio Ruggeri, Jozef Kisel'ák, and Milan Stehlík. "Predicting hourly ozone concentrations using wavelets and ARIMA models." Neural Computing and Applications 31, no. 8 (2019): 4331-4340.

[52] Nazif, Amina, Nurul Izma Mohammed, Amirhossein Malakahmad, and Motasem S. Abualqumboz. "Regression and multivariate models for predicting particulate matter concentration level." Environmental Science and Pollution Research 25, no. 1 (2018): 283-289.

[53] Selvi, S., and M. Chandrasekaran. "Performance evaluation of mathematical predictive modeling for air quality forecasting." Cluster Computing 22, no. 5 (2019): 12481-12493.

[54] Jung, Chau-Ren, Bing-Fang Hwang, and Wei-Ting Chen. "Incorporating long-term satellite-based aerosol optical depth, localized land use data, and meteorological variables to estimate ground-level PM2. 5 concentrations in Taiwan from 2005 to 2015." *Environmental Pollution* **237** (1): 1000-1010, 2018.

[55] Rybarczyk, Yves, and Rasa Zalakeviciute. "Machine learning approaches for outdoor air quality modelling: A systematic review." *Applied Sciences* **8**, no. 12 (17): p.2570, 2019.

[56] Karpatne, Anuj, Imme Ebert-Uphoff, Sai Ravela, Hassan Ali Babaie, and Vipin Kumar. "Machine learning for the geosciences: Challenges and opportunities." *IEEE Transactions on Knowledge and Data Engineering* (18), 2018.

[57] Grange, Stuart K., David C. Carslaw, Alastair C. Lewis, Eirini Boleti, and Christoph Hueglin. "Random forest meteorological normalisation models for Swiss PM 10 trend analysis." *Atmospheric Chemistry and Physics* **18**, no. 9 (19): p.6223-6239, 2019.

[58] Zhao, Fang, and Weide Li. "A Combined Model Based on Feature Selection and WOA for PM2. 5 Concentration Forecasting."*Atmosphere* **10**, no. 4 (20): p.223, 2019.

[59] Isakov, Vlad, Saravanan Arunachalam, Stuart Batterman, Sarah Bereznicki, Janet Burke, Kathie Dionisio, Val Garcia et al. "Air quality modeling in support of the near-road exposures and effects of urban air pollutants study (NEXUS)."*International journal of environmental research and public health* **11**, no. 9 (21): p.8777-8793, 2014.

[60] Zhu, Dixian, Changjie Cai, Tianbao Yang, and Xun Zhou. "A machine learning approach for air quality prediction: Model regularization and optimization."*Big data and cognitive computing* **2**, no. 1 (22): p.5, 2018.

[61] Suleiman, A., M. R. Tight, and A. D. Quinn. "Applying machine learning methods in managing urban concentrations of traffic-related particulate matter (PM10 and PM2. 5)." *Atmospheric Pollution Research* **10**, no. 1 (23): p.134-144, 2018.

[62] Xie, Jingjing, Xiaoxue Wang, Yu Liu, and Yun Bai. "Autoencoder-based deep belief regression network for air particulate matter concentration forecasting." *Journal of Intelligent and Fuzzy Systems* **34**, no. 6 (24): p.3475-3486, 2018.

[63] Soh, Ping-Wei, Kai-Hsiang Chen, Jen-Wei Huang, and Hone-Jay Chu. "Spatial-Temporal pattern analysis and prediction of air quality in Taiwan." *10th International Conference on Ubi-media Computing and Workshops (Ubi-Media)IEEE* ,2017 (2), pp. 1-6.

[64] Bai, Yun, Yong Li, Bo Zeng, Chuan Li, and Jin Zhang. "Hourly PM2. 5 concentration forecast using stacked autoencoder model with emphasis on seasonality." *Journal of Cleaner Production* 224, 2019 : 739-750.

[65] Keller, Joshua P., Casey Olives, Sun-Young Kim, Lianne Sheppard, Paul D. Sampson, Adam A. Szpiro, Assaf P. Oron, Johan Lindström, Sverre Vedal, and Joel D. Kaufman. "A unified spatiotemporal modeling approach for predicting concentrations of multiple air pollutants in the multi-ethnic study of atherosclerosis and air pollution."*Environmental health perspectives*123, no. 4 (2015): 301-309.

[66] Lv, Baolei, Yongtao Hu, Howard H. Chang, Armistead G. Russell, and Yuqi Bai. "Improving the accuracy of daily PM2. 5 distributions derived from the fusion of ground-level measurements with aerosol optical depth observations, a case study in North China."*Environmental science technology*50, no. 9 (2016): 4752-4759.

[67] Fann, Neal, Evan Coffman, Brian Timin, and James T. Kelly. "The estimated change in the level and distribution of PM2. 5-attributable health impacts in the United States: 2005–2014."*Environmental research* 167 (2018): 506-514.

[68] Luo, Zhipeng, Jianqiang Huang, Ke Hu, Xue Li, and Peng Zhang. "AccuAir: Winning solution to air quality prediction for KDD Cup 2018." In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, pp. 1842-1850. 2019.

[69] Sun, Xue, Xiao-San Luo, Jiangbing Xu, Zhen Zhao, Yan Chen, Lichun Wu, Qi Chen, and Dan Zhang. "Spatio-temporal variations and factors of a provincial PM 2.5 pollution in eastern China during 2013–2017 by geostatistics."*Scientific reports* 9, no. 1 (2019): 1-10.

Research candidate

HarishKumar
Research scholar,
Dept. of Computer Science,
Mangalore University.

Research supervisor

Dr. Manjaiah D.H,
Professor,
Dept. of Computer Science,
Mangalore University.